CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 6

Usability of CEFR Companion Volume scales for the development of an analytic rating scale for academic integrated writing assessment

Claudia Harsch, University of Bremen

Valeriia Koval, University of Bremen

Ximena Delgado-Osorio, DIPF | Leibniz Institute for Research and Information in Education

Johannes Hartig, DIPF | Leibniz Institute for Research and Information in Education

https://doi.org/10.37546/JALTSIG.CEFR6-9

This article is open access and licensed under an Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence.

Successful academic writing from sources requires a broad range of competencies. When writing from sources, students are expected to mine source texts for relevant ideas, present these ideas with precision and in necessary depth, have efficient paraphrasing skills and the knowledge of proper source attribution. In order to assess the combination of these skills in writing and to provide diagnostic feedback to the learners, there is a need to design a rating scale where the required skills are operationalized in separate criteria (Knoch 2011). However, this endeavour may be challenging due to the complex nature of the academic integrated writing construct.

This article describes the process of analytic rating scale development in the context of German higher education (HE). We address the issues of construct complexity and the operationalization of the construct elements in rating scale criteria by a combination of theory-based, descriptor-based, empirical, and intuitive approaches to scale development (e.g., Chan, Chihiro and Taylor 2015; Kuiken and Vedder 2021), with a particular focus on the usability of relevant scales from the CEFR Companion Volume (CEFR/CV; Council of Europe 2020). Besides the CEFR scales, we also explore the usability of existing scales for integrated writing and relevant taxonomies (e.g., Keck 2006; Shi 2004). Finally, we present qualitative insights of intuitive expert judgement from a workshop with four content experts who trialled and refined the first draft of the rating scale. The ensuing validation of the rating scale is, however, beyond the scope of this paper and the mixed-methods validation study will be reported elsewhere.

The rating scale development reported here was part of the DFG-funded research project Modelling of academic integrated linguistic competencies, conducted at the University of Bremen and the Leibniz Institute for Research and Information in Education in Frankfurt. The project aim was to evaluate the academic-linguistic preparedness of students taking up English-medium studies in Germany by employing authentic integrated writing tasks and valid assessment procedures. The article offers insights into challenges and critical considerations when developing CEFR-based rating scales for integrated writing, focusing on valid rating criteria, bands, and adapting existing descriptors.

Keywords: rating scale development, CEFR/CV, integrated writing tasks, academic preparedness, validation of a rating scale

1 Introduction

Successful academic writing from sources requires a broad range of competencies. When writing from sources, students are expected to mine source texts for relevant ideas, present these ideas with precision and in necessary depth, and have efficient paraphrasing skills and the knowledge of proper source attribution. In order to assess the combination of these skills in writing and to provide diagnostic feedback to the learners, there is a need to design a rating scale where the required skills are operationalized in separate criteria (Knoch 2011). However, this endeavour may be challenging due to the complex nature of the academic integrated writing construct.

We addressed this challenge in the context of German higher education (HE) by a combination of different approaches to scale development, with a particular focus on exploring how far relevant scales from the CEFR Companion Volume (CEFR/CV; Council of Europe 2020) could be adapted to suit the demands for diagnostic rating scales that aim to foster students' academic writing skills in a low-stakes assessment. Here, we outline how we defined and operationalized relevant construct elements in our rating scale by combining theory-based, descriptor-based, empirical, and intuitive approaches to scale development (e.g., Chan et al. 2015; Kuiken and Vedder 2021). We report detailed analyses of the CEFR/CV scales, other existing rating scales that address integrated writing, as well as relevant taxonomies and models, aiming to offer insights into the feasibility of using the reviewed scales and models for similar rating scale development projects.

2 Background

The study reported here is situated within a larger project examining the dimensionality of integrated academic-linguistic competences. The project was conducted at the University of Bremen and the Leibniz Institute for Research and Information in Education in Germany during 2020-2023 and funded by the German Research Foundation. The project is situated at a crossroads between upper secondary school and university. It aims to assess the academic-linguistic preparedness of school leavers and university freshmen in a context where English as lingua franca is used as medium for instruction (EMI). The expected proficiency in English as a foreign language at this point in education is defined in the national educational standards at B2, with certain aspects reaching C1 of the Common European Framework of Reference (KMK 2014). University language expectations are also expressed via CEFR levels and usually require B2 (sometimes C1) for BA programmes where English is the medium of instruction. Ultimately, the assessment reported here will be used as a low-stakes formative post-entry diagnosis in such study programmes.

We employed integrated reading-into-writing tasks, which have a high level of authenticity in the academic context (Cumming 2013). The tasks were developed by two experienced teachers, one with an EAP background, and the other being an academic faculty member in English teacher education. They designed four integrated reading-into-writing tasks, two of which required students to write a summary, and the other two were opinion tasks where students were asked to argue for or against two possible stances expressed in the source text. Each task contained one continuous source text (approximately 1000 words) taken from introductory textbooks for freshmen in social and natural sciences. We provided detailed instructions regarding how the source text was to be used and what was expected from students. The task development and validation are beyond the scope of this paper and will be reported elsewhere.

The student scripts elicited by the integrated tasks are to be assessed with a diagnostic rating scale that should validly capture salient features of the integrated construct. This construct also considers the intricate relationship between tasks, strategies, and performances, as depicted by the CEFR/CV (2020, p.35). This paper focuses on the development of the rating scale, its horizontal categories and their vertical level description. The quantitative and qualitative validation and the accompanying rater training of the rating scale draft that we report here will be published elsewhere.

3. Diagnostic rating scales for integrated writing tasks

Rating scales have to be fit for their purpose (e.g., Alderson 1991; Knoch, Deygers, and Khamboonruang 2021); our purpose here lies in diagnostic assessment, along with pointing towards future development. Our scale will be used by assessors, and it is intended to be communicated (albeit in a simplified learner-adapted form) with students prior to taking the post-enrolment assessment. Following Knoch (2011), analytic criteria are most suitable for diagnostic assessment, as they allow insights into the different aspects of the targeted construct that are relevant for diagnosing learners' strengths and weaknesses. Hence, we will review relevant literature to define the most salient construct elements for integrated reading-into-writing tasks (summary and argumentative tasks), which will be the basis for our assessment criteria.

A diagnostic rating scale needs enough vertical bands or levels to inform students of strengths and weaknesses and simultaneously imply a prospective route for learner development, i.e., the next higher level on the rating scale. At the same time, raters can only handle a limited number of levels, which should suit the local context (e.g., Myford 2002). Therefore, for our purpose and context, we decided on five levels, ranging from B1, B1+, B2, B2+, to C1, to allow for a range of levels also slightly below and above the targeted level B2 to take up BA studies. This approach is also suggested by the CEFR (2001, particularly section 9.2.2).

The levels of the analytic assessment criteria should be defined by so-called descriptors that qualitatively describe what features are expected at the respective levels (e.g., North 2003). The wording of the descriptors should be informative for assessors (a future adaptation for learners is planned). According to North and Schneider (1998), descriptors should be short, use clear language, be positively worded (wherever possible), describe the levels independently of each other, and not merely use adjectives to differentiate the levels.

In the context of rating integrated reading-into-writing, Cumming (2013) mentions the specific challenge of evaluating the influence of the source text on the writing product. Not only do raters have to detect those ideas that were selected from the source text, raters also need to differentiate between the language produced by learners from that of the source text language, with a particular focus on differentiating verbatim copying, paraphrasing, and language produced independently from the source text. We argue that specific criteria should be dedicated to these aspects in diagnostic rating scales to support raters with these challenging and complex tasks.

4. Approach to rating scale development

The literature reports theory-based, descriptor-based, empirical, and intuitive approaches to rating scale development (e.g., COE 2001; Kuiken and Vedder 2021). In order to develop our integrated construct and hence the horizontal assessment criteria of our rating scale, we first reviewed relevant studies and research that can inform these criteria, thereby relying on a theory-based approach to rating scale development. Next, we needed to describe the vertical levels of the rating scale, i.e., develop the descriptors. For the first draft of our descriptors, we employed all of the aforementioned four approaches.

4.1 Construct and horizontal assessment criteria

Following Knoch (2011), we first examined the theoretical construct underlying the integrated reading-into-writing skills; we reviewed the literature for existing theories, frameworks, and models that can help define the most relevant construct elements, which in turn will constitute our assessment criteria, or in other words the horizontal dimensions of our rating scale. While Knoch and Sitajalabhorn (2013) state that no theory or model of integrated reading-into-writing is available, they list the following construct-relevant elements (Knoch and Sitajalabhorn, 2013, p. 303):

1. Mining/selecting the input text(s) for ideas to be used.
2. Synthesising ideas from various sources or summarising from one source.
3. Transforming the language used in the source text(s).
4. Choosing the organisational structure to be used in writing (which is often different from the structure of the input text).
5. Connecting the ideas in the writing; connecting ideas in the reading with their own ideas.

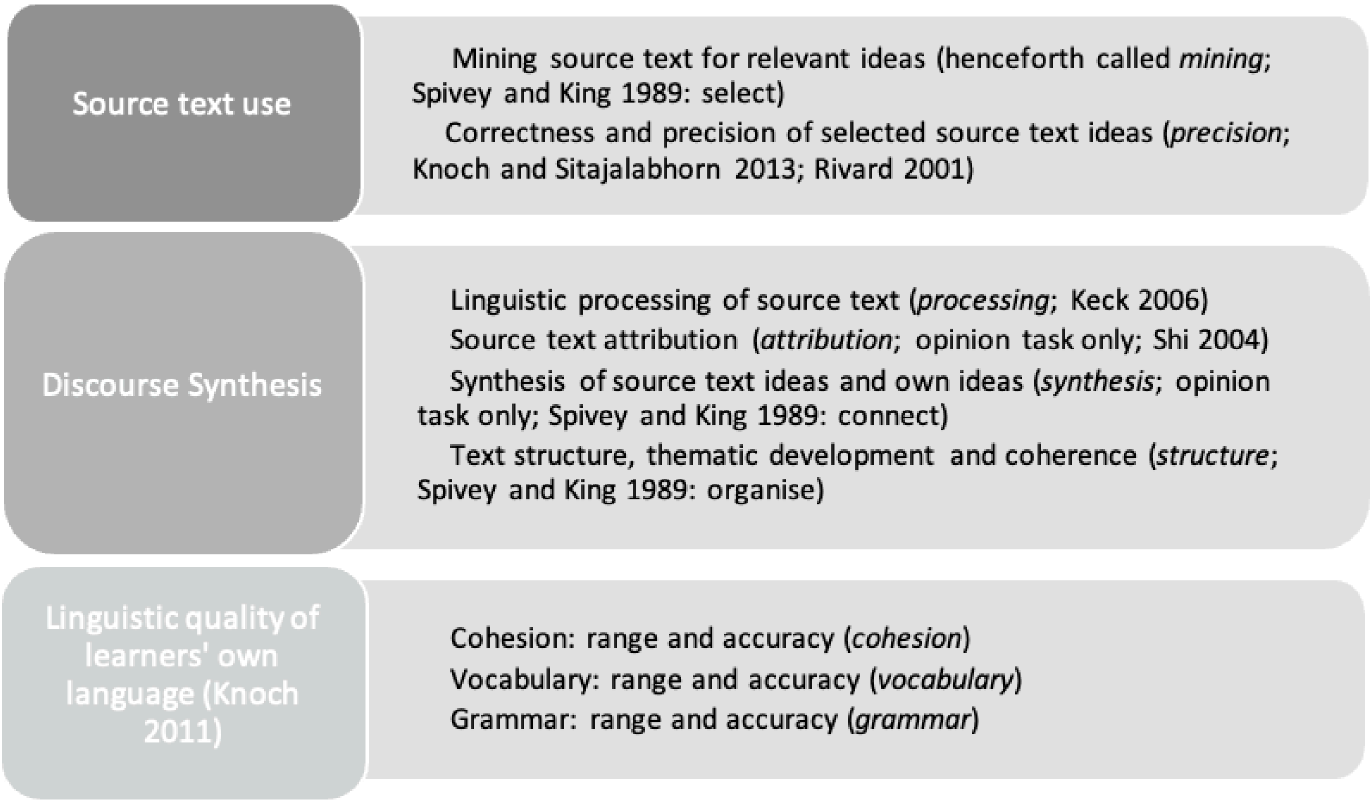
It is apparent that learners need both reading and writing skills (Sawaki et al. 2013), as well as what Spivey and King (1989) called discourse synthesis, i.e., organising the overall structure of one's own writing, considering the structure of the input, selecting relevant ideas from sources, and connecting ideas (from source texts and own ideas). These processes were found more frequently with higher proficiency learners by Plakans (2009) or Plakans and Gebril (2017), showing relevance for the integrated academic writing construct.

Looking at language production and thus the writing part of the construct, Knoch (2011) presents a fairly extensive diagnostic taxonomy, which does, however, not focus on the specifics of integrated writing, such as the accurate presentation of source text ideas (e.g., Knoch and Sitajalabhorn 2013), the quality of the represented ideas (Rivard 2001) or as Li and Wang (2021) called it, the faithfulness with which the ideas from the source text are represented. Moreover, the demand to transform language from the input in order to present ideas from sources in one's own language (e.g., Cumming, 2013) has to be considered. Here, the studies by Keck (2006) on paraphrasing types, and Shi (2004) on textual borrowing and referencing sources can inform the integrated construct, which should include aspects of verbatim borrowing from source texts, the extent and nature of paraphrasing (both semantically and syntactically), and particularly in opinion tasks the element of source text attribution. Shi (2004) demonstrates nicely that task demands can impact the integrated construct and need to be considered, as Knoch and Sitajalabhorn (2013, p. 305) also argue. In our case, we need to particularly consider the demand of the opinion task to develop a coherent line of argument and to present one's stance regarding a particular question raised in the instructions. Finally, regarding the assessment of the quality of students' own language, we employed the three linguistic assessment criteria (i.e., cohesion, vocabulary and grammar, each one subsuming range and accuracy) that are traditionally used for writing assessment in the higher education context that our assessment is situated in.

To sum up, based on the literature reviewed here, we differentiate three broader areas, i.e., source text use, discourse synthesis and the linguistic quality of students' own language. Each area is broken down into several sub-aspects to provide as much diagnostic information as possible. Figure 1 gives an overview of our diagnostic assessment criteria and their main theoretical sources:

For *source text use*, we found the two closely related aspects of selecting the relevant ideas (mining) and of accurately and precisely presenting the selected ideas most relevant (precision). We differentiated these aspects from *discourse synthesis*, as we want to provide diagnostic feedback on reading comprehension, which we believe is best presented via *source text use*. Under *discourse synthesis*, we included the aforementioned aspects of linguistic processing or paraphrasing; for our opinion tasks, we included source text attribution as well as synthesising own and source text ideas; there are no criteria that would only apply to the summary tasks. We also incorporated text structure and thematic development under *discourse synthesis* for both task types, to account for re-organising the source text and -for the opinion task- developing one's own line of argument. Finally, we subsumed the traditional criteria[1] of cohesion, vocabulary and grammar (always with a view to range and accuracy) under *linguistic quality*, thereby shifting the diagnostic focus to the language produced by learners, in order to support raters to differentiate learners' own language from linguistic items borrowed from the source (which is dealt with under *linguistic processing of source text*).

[1] Traditional at least in the higher education context that our assessment is situated in.



Source text use

Mining source text for relevant ideas (henceforth called *mining*; Spivey and King 1989: select)

Correctness and precision of selected source text ideas (*precision*; Knoch and Sitajalabhorn 2013; Rivard 2001)

Discourse Synthesis

Linguistic processing of source text (*processing*; Keck 2006)

Source text attribution (*attribution*; opinion task only; Shi 2004)

Synthesis of source text ideas and own ideas (*synthesis*; opinion task only; Spivey and King 1989: connect)

Text structure, thematic development and coherence (*structure*; Spivey and King 1989: organise)

Linguistic quality of learners' own language (Knoch 2011)

Cohesion: range and accuracy (*cohesion*)

Vocabulary: range and accuracy (*vocabulary*)

Grammar: range and accuracy (*grammar*)

Figure 1. Diagnostic Assessment Criteria.

4.2 Developing rating scale descriptors for the vertical levels

Now that the assessment criteria, i.e., the horizontal dimension of the rating scale, are defined based on a literature review, the next step is to describe the vertical levels of the rating scale for each criterion. As outlined above, our context requires five levels, which we want to derive from or align to the CEFR levels B1 to C1 wherever possible, as this is the frame within which language education in our context is situated. Hence, in a descriptor-based approach, we first analysed the CEFR Companion Volume (CEFR/CV; Council of Europe 2020) for relevant scales and descriptors, before we examined other existing rating scales in the context of (diagnostic) integrated reading-into-writing assessment. We are fully aware that the CEFR/CV scales are proficiency scales and hence need to be specified and adapted to suit our context (see e.g., Alderson 1991). We followed approaches that were established in earlier projects (e.g., Harsch and Martin 2012; Harsch et al. 2020; Rupp et al. 2008). Our aim was to select and, where necessary, adapt existing descriptors to describe our assessment criteria on the five vertical levels that we established as necessary for our diagnostic purpose.

We could, however, not find relevant descriptors for all our assessment criteria and levels, which is why we then resorted to theory-based and empirical approaches: On the one hand, we consulted existing models and insights from research studies (e.g., relevant coding schemes) to help us with formulating missing aspects and descriptors. On the other hand, we employed an empirical approach to qualitatively analysing student scripts, which we had collected in a first trial of the integrated tasks. This served as further source to inform descriptor development, as well as a cross-check whether the main features that we planned to incorporate in the rating scale could actually be found in the scripts, an initial step to validate the scale while still developing it.

A further step in that direction was the intuitive approach that we finally used: With the drafted set of criteria and their descriptors (see Appendix A), we consulted four content experts who reviewed and trialled the draft with selected student scripts; the insights from this consultation were used to revise the draft, the outcome of which is presented in Appendix B. We now describe each of these four approaches in turn.

4.2.1 Descriptor-based approach

We first searched the CEFR/CV for scales relevant to our criteria; we found the nine scales listed in Table 1 below most informative, despite some challenges (see also Harsch et al. 2020), such as no plus (+) levels defined or some defining elements that were not relevant for our context. For example, the CEFR/CV often uses different text types (from simple newspaper articles at B1 to complex academic texts at C1) or different domains (e.g., private life at lower levels to academic or professional domains at higher levels) to differentiate the levels. However, in our context, only the educational domain is relevant, and we only have one source text type (i.e., academic textbooks for freshmen). Hence it was challenging to adapt the descriptors and find differentiating features for the different levels of the rating scale. Overall, we agree with McNamara et al. (2018, p. 25) that the CEFR "is underspecified in terms of the domain of academic literacy", particularly at levels B1/B1+. Table 1 lists the scales that we selected as basis, the abbreviations we used to mark the origin of the descriptors in the scale draft, and the main challenges that we encountered when adapting the descriptors.

Table 1. ***Selected CEFR/CV scales***

Our Criteria	CEFR/CV scales	Abbreviation	Challenges
Mining	PROCESSING TEXT IN WRITING, p.101-103	PT	The construct of a successful summary is not defined in the scale; rather, the levels are differentiated by different source text types, tasks and domains; challenging to adapt for our educational context and academic source texts, and our aim to define summary skills by distinguishing features at different levels.
Precision	READING FOR ORIENTATION, p.55	RFO	The levels are differentiated by different source text types, tasks and domains, which are not always relevant for our context.
Mining	READING FOR INFORMATION AND ARGUMENT, p.56	RFIA	See RFO above, e.g., C1: academic texts vs. B1: newspaper adverts (irrelevant for our context).
Attribution	REPORTS AND ESSAYS, p.68	WRE	See RFO above
Cohesion	COHERENCE AND COHESION, p.141	CC	No cohesion-descriptor at B1+; difficult to define a level between B1: "can link a series of shorter, discrete simple elements…" and B2: "can use a limited number of cohesive devices …".
Vocabulary	VOCABULARY RANGE, p.131	VR	No differentiation between B1 and B1+.
Vocabulary	VOCABULARY CONTROL, p.132	VC	No +levels.
Grammar	GRAMMATICAL ACCURACY, p.132	GA	Descriptors for range of structure not consistent and only mentioned at B1 and B2.
Vocabulary Grammar	ORTHOGRAPHIC CONTROL, p.136	OC	No +levels.

Like in similar scale development projects (e.g., Harsch and Martin 2012, Harsch et al. 2020), we employed a range of adaptation processes, such as splitting or subsuming CEFR/CV descriptors, re-classifying them to fit into our criteria, adding our own wording to specify descriptors for our context or adding missing aspects. Furthermore, we dropped the "can do" wording, as we transformed proficiency scales into rating scales, where the focus is not on what learners, in general can do, but on what raters can observe in text products. We would like to illustrate the different ways of adaptation with three examples. We first list the original CEFR/CV descriptor wording and contrast them with our adaptations in table 2, before we explain the adaptation processes.

Table 2. *Illustration of adaptation processes*

Example	Original wording from CEFR/CV descriptors	Our adaptation[a]
1 subsuming, re-classifying, dropping and/or adding aspects	**RFO B1+:** Can scan longer texts in order to *locate desired information*, and gather information from different parts of a text, or from different texts *in order to fulfil a specific task*. **RFIA B1+:** Can identify the *main conclusions* in clearly signalled argumentative texts. Can recognize the *line of argument* in the treatment of the issue presented, though not necessarily in detail.	Criterion Mining, Level 2/B1+: *Locates* and selects some of the *desired information* (e.g., *main ideas, conclusion, line of argument), in order to fulfil a specific task.*
2 splitting, re-categorizing	**OC B1:** *Spelling, punctuation* and layout are *accurate enough to be followed most of the time*.	Criterion Vocabulary, Level 1/B1 and below: *Spelling* is *accurate enough to be followed most of the time.* Criterion Grammar, Level 1/B1 and below: *Punctuation* is *accurate enough to be followed most of the time.*
3 expanding a concept	**CC B2+:** Can *use a variety of linking words efficiently to mark clearly the relationships between ideas*.	Criterion Cohesion, Level 4/B2+: Uses *a variety of* cohesive devices (e.g. *linking words*, semantic fields) *efficiently to mark clearly the relationships between ideas.*

Note: [a] Text in *italics*/lilac: CEFR/CV wording used in our descriptors; text underlined/in turquoise: our own wording added to CEFR/CV language.

Example 1 illustrates how we subsumed parts of descriptors from different scales (but at the same level) and re-classified them into one criterion (here: mining), thereby dropping irrelevant aspects and adding relevant wording. In example 2, we split one source descriptor and re-categorized two aspects (spelling and punctuation) into two separate criteria, as we subsumed spelling under vocabulary and punctuation under grammar in our scale in order to reduce the number of assessment criteria. Example 3 illustrates how we expanded a concept which we deemed too narrow (here: linking words) to include other aspects (here: semantic fields) that are also relevant for cohesion.

In addition to the scales that we did include, we would also like to list those CEFR/CV scales that we found not useful for our context and purpose. Table 3 gives an overview along with our reasons for exclusion.

Table 3. *Excluded CEFR/CV scales*

CEFR/CV scale	Reasons for exclusion
STRATEGIES TO EXPLAIN A NEW CONCEPT, Subcategory ADAPTING LANGUAGE, p.118	Advantage: integrated focus. Disadvantage: levels differentiated by type of input text (simple to complex texts – not relevant for our context). Uses the operator "to paraphrase" without defining different kinds of paraphrasing (see Keck 2006 or Shi 2004, who have a more relevant approach to defining differing degrees of successful paraphrasing).
STRATEGIES TO SIMPLIFY A TEXT, Subcategory AMPLIFYING A DENSE TEXT, p.121	Levels differentiated by varying domains, target audiences or topics, which is not relevant for our context.
THEMATIC DEVELOPMENT, p.139	The text types used to differentiate the levels are mostly irrelevant for our context; when referring to developing a line of argument, this would only be relevant for the opinion task, but there is no mentioning of the synthesis of source text and own ideas, which forms the basis for argument development in our tasks. For our criterion 4 thematic development, we used a more relevant rating scale that was also based on the CEFR (see below, Rupp et al. 2008).
Written Assessment Grid from Manual, Table C4, p.187ff	Criteria Range and Accuracy: descriptors are very generic and abstract, would need to be specified; moreover, we defined accuracy in relation to range both for our criteria vocabulary and grammar, and thus would have had to re-write all descriptors. Criterion Argument: Levels differentiated by output/genre/text type (e.g., exposition on C1 vs. very brief report on B1), which is not relevant for our context.
Supplementary descriptors from scale ADAPTING LANGUAGE, p.263	Descriptors not consistent, targeting different aspects at each level, which are not relevant for our context.

Since we could not find suitable descriptors in the CEFR/CV for all our criteria and levels, we resorted to other existing rating scales that focus on integrated writing, diagnostic assessment or are based on the CEFR. Table 4 lists the two scales that we used and gives our reasons.

Table 4. *Additional scales that we included*

Our Criteria	Source Scale	Abbreviation	Reasons for selection
Vocabulary, Grammar	Pearson (2015) Global Scale of English, scale WRITTEN PRODUCTION: criteria range and accuracy, (pp.5-6).	GSE	GSE based on CEFR, targeting academic domain, all +levels defined; we used it to describe the missing +levels in CEFR/CV scales Vocabulary Control and Orthographic Control for our criteria vocabulary and grammar.
Structure, Cohesion, Grammar	IQB-Scales (Rupp et al., 2008): RATING SCALES FOR WRITING TASKS, levels B1-C1, criteria organization and grammar, (pp.149-155).	IQB	CEFR-based rating scale, validated (Harsch and Martin 2012); even if no +levels are defined and it is not targeting integrated writing, we found the specifications and adaptations suitable for our purposes, particularly the approach to set parts of descriptors in *italics* to mark their nature as rating guidelines (e.g., error treatment, to prevent raters from looking for errors, see *italics* in Appendix A). We used some of the wording for our criteria structure, cohesion and grammar.

There were four other scales that we consulted and analysed, but found less suitable for various reasons: One was the IELTS (2013) Writing Band Descriptors for Task 1 (public version). The IELTS academic task 1 requires a summary of a discontinuous text, which is of less relevance for our context, as is the criterion task achievement; the nine band descriptors do not address paraphrasing or textual borrowing. The descriptors of the linguistic criteria are not aligned to the CEFR; they describe a range of very limited proficiency seemingly below B1 requirements ("can only use a few isolated words; cannot use sentence form at all") to a high level of proficiency, where the bands are often differentiated by adjectives such as "extremely limited" vs. "very limited".

The second scale we consulted was the TOEFL Integrated Writing Rubrics (ETS n.d.). The TOEFL integrated task requires students to use input from listening and reading sources to fulfil a specified task. The holistic scale describes five bands that are not aligned to the CEFR. The scale covers relevant aspects such as selection and accuracy of source ideas, coherence and organization; yet linguistic aspects are defined by the presence or absence of errors. Paraphrasing or textual borrowing is not sufficiently addressed. The lower two levels seem to describe performance below CEFR B1 requirements.

We then analysed the Integrated Skills of English ISE III Task 3—Reading into Writing Rating Scale (Trinity College London n.d.). The ISE III integrated task requires test takers to collate relevant information from several shorter reading texts to fulfil a specified writing task. The rating scale differentiates reading/writing aspects on the one hand, and task fulfilment on the other on four bands. The bands are not aligned to the CEFR, and they are mainly differentiated by the adjectives "excellent", "good", "acceptable", and "poor". Moreover, summary and paraphrasing skills are not sufficiently defined; only level 1 ("heavy lifting and many disconnected ideas") and level 3 ("very limited lifting and few disconnected ideas") add information beyond "poor" respectively "good" summary/paraphrasing skills. We assume that such a differentiation will not sufficiently support raters to differentiate paraphrasing/summarizing skills on our five targeted levels.

Finally, we checked the CUNY Assessment Test in Writing Analytic Scoring Rubric (CUNY 2012, p.4). While the reading-into-writing assessment has a comparable purpose (low stakes, freshmen), and a comparable opinion task, it uses a much shorter reading text (250-300 words). The five analytic criteria cover similar aspects, yet these aspects are grouped very differently to our criteria; e.g., understanding of input ideas, integrating them with own ideas and responding to input are grouped in the first criterion. While the different aspects are coherently defined on six levels, the levels are not aligned with the CEFR. The two lowest levels target proficiency below B1, while the highest level perhaps reaches above C1.

4.2.2. Theory-based approach

Based on the extensive scale- and descriptor-analyses reported above, we did not find sufficiently precise descriptors in the CEFR/CV or other existing scales, particularly for our criteria in the dimension *discourse synthesis*. Here, we resorted to a theory-based approach and selected taxonomies or coding schemes from relevant research projects as basis to formulate our own descriptors. Table 6 gives an overview of the sources used.

Table 6. *Additional sources*

Our Criteria	Source	Details and comments
Processing	Keck (2006)	We used the taxonomy "near copy, minimal revision, moderate revision, substantial revision" (p. 268) to formulate descriptors regarding the aspect of paraphrasing/textual borrowing.
Processing	Shi (2004)	We used the coding scheme "exact copy, slightly modified, modified" (p. 196) to formulate descriptors regarding the aspect of paraphrasing/textual borrowing.
Attribution	Shi (2004)	We used the coding scheme "with referencing, without referencing" (p. 196) to formulate descriptors regarding the aspect of attribution of ideas.
Structure	Li (2014)	We employed the aspect of "logically rearranging" ideas in one's own text (p.13) in some descriptors in our text structure criterion.

The exact adaptations are referenced and colour-coded in our rating scale draft 1 in Appendix A.

4.2.3. Empirical approach

For piloting the four integrated writing tasks, we invited freshmen in programmes with English as medium of instruction as well as participants in our English for academic purposes course at the languages centre. The freshmen enter the university with at least CEFR level B2 in English, and this is also the level required for the language course. 84 students volunteered to participate. Each student had one hour to work on one of the four tasks. We expected about 300 words output for the summary tasks and 350 words for the opinion tasks; this was stated in the task instructions.

We analysed the 84 collected scripts (between 20 and 22 per task) with regard to seminal features that we used to define the rating scale criteria. The project team first sorted the scripts intuitively into low/medium/high proficient scripts before analysing them in more detail. The analyses happened around the time of the expert workshop (see 4.2.4 below), with some analyses taking place before, and particularly the analyses regarding the selection of relevant source text (ST) ideas and the precision with which they were presented taking place after the expert workshop. Here, we report a synopsis of our analyses.

We analysed all 84 scripts for task-dependent features, such as selecting relevant ST ideas and attributing them, or using and integrating own ideas in the opinion task. Regarding our criterion Mining, we analysed the scripts against the list of relevant ideas that was developed as rating guide (see 4.2.4 below). We found that all ideas that we marked as relevant were used, some by all students, others less frequently; in cases where only a minority of students had selected a specific ST idea, we revised the list.

For the opinion task, we examined the 42 scripts with regard to students attributing selected ideas to the ST, which we found more with scripts at the higher end, while scripts in the low-proficiency pile did not attribute ideas. We also found that about 50% of the scripts in the opinion tasks included own ideas; therefore, we developed descriptors addressing this feature. Some students used only ideas from the ST to support their stance, others used mainly their own ideas, yet others used a balanced approach (these tended to be the more proficient ones). We also analysed the macro-structure in these scripts and found three main approaches to developing one's stance: students either argued for or against one of the two positions in the ST or came to a balanced stance. The approaches seemed unrelated to the high- or low-proficiency piles into which we had sorted the scripts; hence we allowed all possible stances as equally valuable, as long as the student's stance became apparent and was well-informed.

We present the initial rating scale draft in Appendix A, where we colour-coded and referenced all sources for the descriptors, using the abbreviations listed in the tables above, to indicate the exact source of the wording we borrowed from existing descriptors, derived from theoretical models and coding schemes, or based on student script analyses. Our own wording that we used to adapt the descriptors for consistency and appropriacy for our context and purpose is kept unmarked in black. Table 7 lists all sources that we used as basis for our descriptor-wording:

Table 7. *Sources of descriptor-wording*

Criterion	1a Mining ST	1b ST ideas Correctness	2 Linguistic processing	3a ST attribution	3b Synthesis ST own ideas	4 Text structure,	5 Cohesion	6 Vocab	7 Grammar
Descriptor sources	- CEFR/ CV	- scripts	- Shi 2004 - Keck 2006 - scripts	*- CEFR/CV* *- scripts*	*- scripts*	- CEFR/CV - IQB - Li, 2014	- CEFR/ CV - IQB	- CEFR/ CV - GSE	-CEFR/CV - IQB - GSE

Note: Criteria 3a and 3b apply only to the opinion tasks.

Despite all efforts, there are a few empty cells in the matrix in Appendix A, as we did not manage to develop suitable descriptors for all levels. We still had the intention to fill these either in the expert workshop or later during rater training.

4.2.4. Intuitive approach

With this first draft of the rating scale, we conducted a two-day workshop with the two experts who had developed the integrated tasks, and two experienced teachers of English for academic purposes. The experts were first familiarised with the tasks and the rating scale draft. Then, they were provided with three scripts per task and asked to evaluate the scripts using the criteria for the dimensions *Source Text Use* and *Discourse Synthesis*. We discussed results, digressions, justifications, as well as ways to improve the rating scale. We protocolled the discussions and outcomes. The findings reported here are based on the protocol and focus only on feedback for the rating scale.

Overall, the experts found the criteria meaningful and relevant, and the five levels feasible. With regard to the criterion *Mining*, they recommended developing the aforementioned list of relevant ST ideas, in order to better support the raters. Hence, we developed task-specific lists in the workshop, spelling out for the summary tasks which main ideas we expected to be included, and for the opinion tasks which ideas we regarded as relevant (from which writers were expected to choose a few, depending on their stance). With regard to differentiating levels 4 and 5, the experts suggested adding ”may contain some irrelevant ideas“ for Level 4. They also suggested adding a statement on the depth of understanding of the ST ideas for the higher levels. Criterion *Precision* was perceived as helpful and easy to apply, but the experts suggested to add a qualification for level 5, to specify that here a high level of precision of the selected ideas is expected.

With regard to the criterion *Processing*, the experts found it difficult to distinguish ST wording from students' own wording, and recommended further support for the raters. This recommendation coincided with the development of an automated tool to highlight (strings of) words copied from the ST, specifically designed for our project by the research group of Prof. Zesch, then University Duisburg-Essen, Germany. Another recommendation was to add a special code for cases where writers only used their own ideas (and hence no paraphrasing could occur). The criterion *Attribution* was perceived as clearly worded and feasible, while for criterion *Synthesis*, the experts recommended specifying that the writer's stance needs to be related to the ST, the presented ideas (ST and own) need to be relevant for the stance, the ST ideas and own ideas need to be meaningfully related to each other, as well as well-informed at the highest level. Finally, for criterion *Structure*, the experts recommended to add the expectation for the highest level that a logical development is expected not only for the text as a whole, but also on the paragraph level, and to use this feature for the gradation on the lower levels.

We used these recommendations to revise the scale, and we present the revised draft 2 in Appendix B, where we highlight all changes to draft 1.

5 Discussion and conclusions

We found the definition of relevant construct elements and their categorisation into assessment criteria challenging, yet manageable; the research literature provides a sufficient basis upon which to define relevant construct elements, and when taking the local context into account, a feasible solution to categorising these elements into assessment criteria could be developed. However, finding suitable descriptors to describe these criteria proved to be more challenging. While the CEFR/CV provides a rich source of scales and descriptions, not all scales and descriptors were feasible for our context and construct elements. This holds particularly true for those scales that use domains, target audiences or topics irrelevant to our context. In addition, other CV scales showed inconsistencies regarding the features that are described, or the wording with which these features are graded across the different scale levels. Hence, in the majority of cases, we needed to select and adapt the existing CV descriptors, mainly by splitting existing descriptors into separate criteria, subsuming different descriptors under one criterion, re-categorising certain aspects to suit our criteria, dropping certain aspects from existing descriptors, or expanding certain concepts to entail all relevant construct elements. These adaptations, which chime with Harsch and Marin (2012) or Harsch et al. (2020), were not only necessary for the CEFR/CV scales, but also necessary for the other existing scales and taxonomies that we used.

A major issue with other existing scales occurred when descriptors defined the construct by itself, e.g., when paraphrasing was defined by having good paraphrasing skills, which happened in a surprising number of instances. Another recurring problem was when scale levels were differentiated solely by verbal gradations, such as 'poor – acceptable – good'. We also dropped scales that were not aligned to the CEFR as we would have needed a further step of aligning existing descriptors to CEFR levels.

Ultimately, as we did not find sufficient and suitable CEFR/CV descriptors for our criteria targeting *source text use* and *discourse synthesis*, we do not claim CEFR alignment for these dimensions. Here, we found other existing scales and taxonomies a useful and helpful addition. Equally, we recommend a combination of all available approaches to scale development, be it intuitive, empirical, descriptor- or theory-based, in order to capture relevant elements and features from all possible angles.

The next step was to validate the thus developed rating scale, which we addressed in a combination of scale trialling and rater training (as recommended by Harsch and Martin 2012), in order to revise the scale descriptors based on empirical rating data (reported elsewhere). We then can validate with students whether the information gained by the analytic rating scale yields meaningful diagnostic feedback.

6 Acknowledgement

Our thanks goes to the task developers and participants in the expert workshop, as well as to all students who participated in our trial. Furthermore, we would like to thank our student assistants Helena Gaubitz, Larissa Junker and Annika Lamers, University of Bremen, for contributing to the qualitative script analyses. Our heartfelt thanks goes to our colleague Heike Neumann, Concordia University, Montréal, Canada for her careful reading of an early version of the manuscript and for her insightful comments and suggestions that improved the manuscript greatly. All remaining errors are ours.

7 References

Alderson, Charles. 1991. Bands and scores. In Charles Alderson and Brian North (Eds.), *Language testing in the 1990s* (pp. 71–86). London: Macmillan.

Chan, Sathena, Chihiro Inoue and Linda Taylor. 2015. Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, 26, 20-37.

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

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg: Language Policy Division. Available at https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4 (accessed 17 May 2023).

Cumming, Alastair. 2013. Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, 10(1), 1–8.

CUNY. 2012. *CUNY Assessment Test in Writing (CATW)*. Available at https://www.cuny.edu/wp-content/uploads/sites/4/page-assets/academics/testing/CATWInformationforStudentsandpracticeweb.pdf (accessed 1 September 2022).

ETS. Not dated. *TOEFL Writing Rubrics*. Available at https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf (accessed 1 September 2022).

Harsch, Claudia, de la Caridad Collada Peña, Ivone, Gutiérrez Baffil, Tamara, Castro Álvarez, Pedro and Ioani García Fernández. 2020. Interpretation of the CEFR Companion Volume for developing rating scales in Cuban higher education. *CEFR Journal Research and Practice 3*, 87-97. https://doi.org/10.37546/JALTSIG.CEFR3-5.

Harsch, Claudia and Guido Martin. 2012. Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. In *Assessing Writing* 17, 228–250.

IELTS. 2013. *IELTS TASK 1 Writing band descriptors* (public version). Available at https://www.ielts.org/-/media/pdfs/writing-band-descriptors-task-1.ashx?la=en (accessed 1 September 2022).

KMK. 2014. *Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife*. Köln: Wolters.

Keck, Casey. 2006. The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing,* 15. 261–278.

Knoch, Ute. 2011. Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, 16. 81-96. 10.1016/j.asw.2011.02.003.

Knoch, Ute, Bart Deygers and Apichat Khamboonruang. 2021. Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing, 38*(4), 602-626. https://doi.org/10.1177/0265532221994052.

Knoch, Ute and Woranon Sitajalabhorn. 2013. A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes. *Assessing Writing*, 18(4), 300–308.

Kuiken, Folkert and Ineke Vedder. 2021. Scoring Approaches: Scales/Rubrics. In Paula Winke and Tineke Brunfaut (Eds.), *The Routledge Handbook of Second Language Testing* (1st ed.). Routledge. https://doi.org/10.4324/9781351034784.

Li, Jiuliang. 2014. The role of reading and writing in summarization as an integrated task. *Language Testing in Asia* 4:3.

Li, Jiuliang and Qian Wang. 2021. Development and validation of a rating scale for summarization as an integrated task. *Asian-Pacific Journal of Second and Foreign Language Education 6*(11). https://doi.org/10.1186/s40862-021-00113-6.

McNamara, Tim, Janne Morton, Neomy Storch and Celia Thompson. 2018. Students' Accounts of Their First-Year Undergraduate Academic Writing Experience: Implications for the Use of the CEFR, *Language Assessment Quarterly, 15*:1, 16-28, https://10.1080/15434303.2017.140542.

Myford, Carol M. 2002. Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education*, 15 (2), 187–215.

North, Brian. 2003. *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats*. TOEFL Monograph, 24.

North, Brian and Günther Schneider. 1998. Scaling descriptors for language proficiency scales. Language Testing, 15 (2), 217–263.

Pearson Education. 2015. *Global Scale of English Learning Objectives for Academic English*. Available at http://www.b-li.ir/ar/06CEFR/4.GSE_LO_Academic_English.pdf (accessed 12 February 2020).

Plakans, Lia. 2009. Discourse synthesis in integrated second language assessment. *Language Testing*, 26(4), 561–587.

Plakans, Lia and Atta Gebril. 2017. Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing 31*, 98-112.

Rivard, Leonard P. 2001. Summary Writing: A Multi-Grade Study of French-Immersion and Francophone Secondary Students. *Language, Culture and Curriculum, 14*(2), 171-186, https://doi.org/10.1080/07908310108666620.

Rupp, Andree A., Miriam Vock, Claudia Harsch and Olaf Köller. 2008. *Developing standards-based assessment items for English as a first foreign language – Context, processes, and outcomes in Germany* (Bd. 1). Münster: Waxmann.

Sawaki, Yasuyo, Thomas Quinlan and Yong-Won Lee. 2013. Understanding Learner Strengths and Weaknesses: Assessing Performance on an Integrated Writing Task, *Language Assessment Quarterly*, 10:1, 73-95, DOI https://doi.org/ 10.1080/15434303.2011.633305

Shi, Ling. 2004. Textual borrowing in second language writing. *Written Communication,* 21. 171–200.

Spivey, Nancy. N. and James R. King. 1989. Readers as writers composing from sources. *Reading Research Quarterly*, 24, 7-26.

Trinity College London (n.d.). ISE III Task 3 Reading into writing rating scale. Available at https://www.trinitycollege.com/resource/?id=7468 (accessed 1 September 2022).

8 Biographies

Claudia Harsch (University of Bremen) is Professor for Research in Language Learning and Teaching at the Faculty for Languages and Literature at the University of Bremen since 2015. She is also the Director of the Languages Centre of the Universities in the Land Bremen. Her research focuses on language assessment and language learning/teaching in higher education contexts.

Valeriia Koval (University of Bremen) is a PhD candidate and a research associate at the University of Bremen since 2020. She holds a Master‘s Degree in applied linguistics from the University of Bonn, Germany. Her research focus is assessment of academic integrated writing and rater cognition.

Ximena Delgado-Osorio (DIPF | Leibniz Institute for Research and Information in Education) is a PhD candidate and a research associate in the Educational Measurement unit at the DIPF | Leibniz Institute for Research and Information in Education since 2020. She studied psychology at the University of the Andes (Colombia) and obtained her Master’s Degree in Psychology: Learning Sciences from the University of Munich.

Johannes Hartig (DIPF | Leibniz Institute for Research and Information in Education) is Professor of Educational Measurement and head of the Educational Measurement unit in the department of Teacher and Teaching Quality at the DIPF | Leibniz Institute for Research and Information in Education since 2010. His research focus lies on modelling and measuring educational outcomes and effects of instruction.

Appendix A: Rating scale Draft1

Sources of descriptors: MASK project team, CEFR, analysis of pilot texts, Shi, 2004, Keck, 2006, Li, 2014, IQB descriptors, Pearson GSE

	Source text ST use		*Discourse synthesis* *(Attribution and Synthesis for opinion task only)*				*Linguistic quality*		
Level	**Mining ST for relevant ideas**	**Precision ST ideas**	**Linguistic processing ST**	**Attribution ST**	**Synthesis ST – own ideas**	**Text structure, them. development**	**Cohesion**	**Vocabulary range & accuracy**	**Grammar range & accuracy**
5 | C1 and above	All relevant main ideas selected No irrelevant details or own ideas (summary).	All ST ideas are presented correctly.	Substantial revision: Expresses all ST ideas in own words (only key words are used with quotation marks). Reformulates syntax of ST.	Clear distinction between own and ST ideas (WRE: C2). All ideas taken from source text are appropriately attributed.	Takes a clear stance, meaningfully relating ST ideas and own ideas to task at hand.	Macrostructure clear/appropriate for task. Rearranges ST elements (and if apl. own ideas) into logical order (not necessarily that of ST). Appropriate paragraphs.	Shows consistent and continuous controlled use of a repertoire of cohesive devices (e.g. referencing, semantic fields, connectors) on sentence and paragraph levels, which contributes to the coherence of the text.	Broad lexical repertoire allowing gaps to be readily overcome with circumlocutions. Good command of common idiomatic expressions and colloquialisms (VR: C1). Occasional minor slips but no significant vocabulary (VC: C1) or spelling errors (VC: C1).	Broad repertoire of linguistic structures and complex sentence patterns. Consistently maintains a high degree of grammatical accuracy (including complex structures). Errors are rare and difficult to spot. Punctuation is consistent and helpful (OC: C1).
4 | B2+	Identifies (RFO: B2+) and selects the majority of relevant and useful (RFO: B2+) ideas of particular	Majority of ST ideas are presented correctly.	*[no descriptors available]*	*[no descriptors available]*	*[no descriptors available]*	*[no descriptors available]*	Uses a variety of (CC: B2+) cohesive devices (e.g. linking words (CC: B2+), semantic fields) efficiently to mark clearly the	Good and varied range of vocabulary and collocations. Is able to express task-relevant ideas and if appl. Opinions.	Good grammatical (GA: B2+) range and control. *Occasional 'slips' or non-systematic errors and minor flaws in sentence structure may*

	sections for the task at hand (RFO: B2+).						relationships between ideas (CC: B2+).		*occur, but they are rare (GA: B2+).* Very few mistakes in punctuation.
3 | **B2**	Identifies and selects most of the relevant content (e.g. contrasting arguments, problem-solution presentation, cause-effect relationships) RFIA: B2) *There may be some irrelevant details from ST.*	Most ideas are presented correctly. *There may be some (minor) misinterpretations.*	Moderate revision: Paraphrases majority of ST ideas (There may be occasional use of ST strings of words that are only slightly modified by adding/ deleting words or using synonyms for content words.) Reformulates majority of syntactical structures.	Overall manages to distinguish between own and ST ideas. *Some ST ideas may not be appropriately attributed*.	Takes a stance and on the whole manages to relate own ideas meaningfully to ST ideas and task. *Own ideas may only be partially relevant.*	Macrostructure on the whole clearly developed, *although there may be some 'jumpiness.* Attempts to rearrange ST elements (and if apl. own ideas) into a logical order, *though not fully successful*. Paragraphs mostly logical. *Appropriate thematic development may compensate for missing paragraphs.*	Uses a limited number of cohesive devices to link his/her utterances into clear, coherent discourse (CC: B2). *Errors in the field of cohesive devices may occur occasionally but do not impede understanding.*	Good range of vocabulary and collocations (VR: B2). Attempts to vary formulation to avoid frequent repetition (VR: B2), *though not always successful*. Accuracy is generally high, *though some incorrect word choice may occur without hindering communication (VC: B2).* Spelling is reasonably accurate *but may show signs of mother tongue influence (OC: B2).*	Good range of also infrequent structures and some complex sentence patterns. Shows a relatively high degree of grammatical control (GA: B2). *May use complex structures rigidly with some inaccuracy. Does not make* impeding *errors (GA: B2).* Punctuation is reasonably accurate *but may show signs of mother tongue influence (OC: B2).*

	Source text ST use		*Discourse synthesis* *(Attribution and Synthesis for opinion task only)*				*Linguistic quality*		

Level	Mining ST for relevant ideas	Precision ST ideas	Linguistic processing ST	Attribution ST	Synthesis ST – own ideas	Text structure, them. development	Cohesion	Vocabulary range & accuracy	Grammar range & accuracy
2 | B1+	Locates and selects some of the desired information (RFO: B1+) (e.g. main ideas, conclusion, line of argument), (RFIA: B1+) in order to fulfil a specific task (RFO: B1+). Includes some irrelevant details.	*Some of the ideas may be interpreted incorrectly* .	Minimal revision: Attempts to paraphrase but not always successful (e.g. Strings of words slightly modified by adding/ deleting words or using synonyms for content words). ST phrases are usually not referenced. Reformulates some syntactical structures.	*[no descriptors available]*	*May take a stance* but only partially manages to relate own ideas to ST ideas and task (e.g. does not provide reasoning to support stance). *Own ideas may not all be relevant or meaningfully related to ST ideas/task.*	*[no descriptors available]*	*[no descriptors available]*	Sufficient range of vocabulary (VR: B1). Some repetitive use of vocabulary. *May make mistakes in spelling of less familiar words.*	Good range of frequent structures. Generally good control *though mother tongue influence (GA: B1+) may be noticeable.* *Errors may occur, but it is clear what he/she is trying to express (GA: B1+).*
1 | B1 and below	Recognizes the most significant	*Majority of selected ideas may be*	Near copy: Major instances of lifting from	Generally difficult *for reader to distinguish*	Barely relies on ST for argumentation, offering own (mis-)	Orders a series of shorter discrete elements into a	Links discrete elements using a limited number of	Sufficient range of vocabulary, with some circumlocutions	Uses a repertoire of frequently used "routines" and patterns associated with more

	points (RFIA: B1). Selects only a minority of the relevant main ideas. Includes irrelevant details or irrelevant own ideas (summary).	*misinterpreted.*	ST (usually without referencing).	*between ST and own ideas.* Ideas from source text are generally not attributed to ST.	interpretation of the topic. Or merely summarizes ST, not adding relevant own stance.	linear sequence of points. Structure/them. development follows ST; Or: *may lack a logical order appropriate for the task.* Paragraphs usually not appropriate (if used at all).	cohesive devices (CC: B1). Shows reasonable control of common cohesive devices *but may overuse certain devices or show a mechanical use.* *The use of more elaborate cohesive devices may sometimes impede communication.*	(VR: B1), repetitions. *May show some instances of inappropriate vocabulary use.* *Major errors may occur when expressing more complex thoughts (VC: B1).* Spelling is accurate enough to be followed most of the time (OC: B1).	predictable situations reasonably accurately (GA: B1). *May attempt complex patterns but generally unsuccessfully.* Punctuation is accurate enough to be followed most of the time (OC: B1).

Appendix B: Rating Scale draft2 after Expert Workshop. All changes to draft1 are marked in red.

	Source text ST use (Reading for relevant main ideas, deep vs. superficial understanding)		*Discourse synthesis (meaning making process)* *(Attribution and Synthesis for opinion task only)*				*Linguistic quality (of the writer's own words)*		
Level	**Mining ST for relevant ideas**	**Precision ST ideas**	**2. Linguistic processing ST**	**Attribution ST**	**Synthesis ST – own ideas**	**Text structure, them. development/coherence**	**Cohesion (within/across sentences)**	**Vocabulary range & accuracy**	**Grammar range& accuracy**
5 | C1 and above	-All relevant main ideas selected, presented in necessary depth -No irrelevant details or own ideas (summary). (Deep understanding)	All ST ideas are presented correctly and precisely.	Substantial revision: -Expresses all ST ideas in own words (only key words are used with quotation marks). -Reformulates syntax of ST.	-Clear distinction between own and ST ideas. -All ideas taken from source text are appropriately attributed.	-Takes a clear stance with a well-informed opinion, meaningfully relating ST ideas and own ideas to task at hand. -Bases argumentation on relevant ST ideas throughout the text.	-Macrostructure clear/appropriate for task. -Rearranges ST elements (and if apl. own ideas) into logical order (not necessarily that of ST). -Appropriate paragraphs that are logical in themselves.	-Shows consistent and continuous controlled use of a repertoire of cohesive devices (e.g. referencing, semantic fields, connectors) on sentence and paragraph levels, which contributes to the coherence of the text.	-Broad lexical repertoire allowing gaps to be readily overcome with circumlocutions. -Good command of common idiomatic expressions and colloquialisms. -Occasional minor slips but no significant vocabulary or spelling errors.	-Broad repertoire of linguistic structures and complex sentence patterns. -Consistently maintains a high degree of grammatical accuracy (including complex structures). Errors are rare and difficult to spot. -Punctuation is consistent and helpful.
4 | B2+	-All relevant and useful ideas selected but not all in necessary depth, or	More than 3, but not yet enough for 5	More than 3, but not yet enough for 5	More than 3, but not yet enough for 5	More than 3, but not yet enough for 5	More than 3, but not yet enough for 5	-Uses a variety of cohesive devices (e.g. linking words, semantic fields) efficiently to	-Good and varied range of vocabulary and collocations. - Is able to express task-relevant ideas and if appl. opinions.	-Good grammatical range and control. *-Occasional 'slips' or non-systematic errors and minor flaws in sentence structure may*

	- some irrelevant details						mark clearly the relationships between ideas.		*occur, but they are rare.* -Very few mistakes in punctuation.
3 | B2	-Majority of the relevant content selected, but not necessarily all in required depth, *(differentiation between main ideas and irrelevant details not yet fully consistent)* *- There may be some irrelevant details.*	-Most ideas are presented correctly. *-There may be some (minor) misinterpretations or imprecisions.*	Moderate revision: -Paraphrases majority of ST ideas (*There may be occasional use of ST strings of words that are only slightly modified by adding/ deleting words or using synonyms for content words.*) -Reformulates majority of syntactical structures.	-Overall manages to distinguish between own and ST ideas. *-Some ST ideas may not be appropriately attributed.*	-Takes a (more or less informed) stance and on the whole manages to relate own ideas meaningfully to ST ideas and task. *-Own ideas may not always be relevant.* *-Argumentation may not fully be based on ST ideas.*	-Macrostructure on the whole clearly developed, *although there may be some 'jumpiness'.* -Attempts to rearrang ST elements (and if apl. own ideas) into a logical order, *though not fully successful.* -Paragraphs mostly logical. *Appropriate thematic development may compensate for missing (or illogically developed) paragraphs.*	-Uses a limited number of cohesive devices to link his/her utterances into clear, coherent discourse. *-Errors in the field of cohesive devices may occur occasionally but do not impede understanding.*	-Good range of vocabulary and collocations. -Attempts to vary formulation to avoid frequent repetition, *though not always successful.* -Accuracy is generally high, *though some incorrect word choice may occur without hindering communication.* -Spelling is reasonably accurate *but may show signs of mother tongue influence.*	-Good range of also infrequent structures and some complex sentence patterns. -Shows a relatively high degree of grammatical control. *-May use complex structures rigidly with some inaccuracy. Does not make impeding errors.* -Punctuation is reasonably accurate *but may show signs of mother tongue influence.*
	Source text ST use (Reading for relevant main ideas, deep vs. superficial understanding)		*Discourse synthesis (meaning making process)* *(Attribution and Synthesis for opinion task only)*				*Linguistic quality (of the writer's own words)*		

	Mining ST for relevant ideas	**Precision ST ideas**	**2. Linguistic processing ST**	**Attribution ST**	**Synthesis ST – own ideas**	**Text structure, them. development/coherence**	**Cohesion (within/across sentences)**	**Vocabulary range & accuracy**	**Grammar range & accuracy**
2 | B1+	- Some of the desired information selected, not necessarily in required depth. -Includes some irrelevant details.	-*Some of the ideas may be interpreted incorrectly.* -*Ideas presented with some imprecision.*	Minimal revision: -Attempts to paraphrase but not always successful (e.g. Strings of words slightly modified by adding/ deleting words or using synonyms for content words). -Reformulates some syntactical structures.	More than 1, but not yet enough for 3	-*May take a stance but opinion is not informed,* only partially manages to relate own ideas to ST ideas and task *(e.g. does not provide reasoning to support stance or only partially bases argumentation on ST ideas).* -*Own ideas may not all be relevant or meaningfully related to ST ideas/task.*	-Identifiable attempt at macrostructure, but not fully successful *(e.g. intro & conclusion but no appropriate middle part).* -Attempt at paragraphs that may not always be logical.	More than B1 but not yet enough for B2	-Sufficient range of vocabulary. Some repetitive use of vocabulary. -*May make mistakes in spelling of less familiar words.*	-Good range of frequent structures. -Generally good control *though mother tongue influence may be noticeable.* -*Errors may occur, but it is clear what he/she is trying to express.*
1 | B1 and below	-Only the most significant points/a minority of the relevant	*Majority of selected ideas may be misinterpreted.*	Near copy: -Major instances of lifting from ST (usually	-Generally difficult for reader to distinguish between	-No clear stance. -Barely relies on ST for argumentation, offering own (mis-)	-Orders a series of shorter discrete elements into a linear sequence of points. Structure/thematic development follows ST, although not	-Links discrete elements using a limited number of cohesive devices.	-Sufficient range of vocabulary, with some circumlocutions, repetitions. -*May show some instances of*	-Uses a repertoire of frequently used "routines" and patterns associated with more predictable situations

	main ideas selected. -Includes irrelevant details or irrelevant own ideas (summary).		without referencing). Add Code 0 (not applicable): writes only own ideas	ST and own ideas. -Ideas from source text are generally not attributed to ST.	interpretation of the topic. -Or merely summarizes ST, not adding relevant own stance.	appropriate for the task. Or: *may lack a logical order appropriate for the task.* -Paragraphs usually not appropriate (if used at all).	-Shows reasonable control of common cohesive devices *but may overuse certain devices or show a mechanical use.* -*The use of more elaborate cohesive devices may sometimes impede communication*.	*inappropriate vocabulary use.* *Major errors may occur when expressing more complex thoughts.* -Spelling is accurate enough to be followed most of the time.	reasonably accurately. -*May attempt complex patterns but generally unsuccessfully.* -Punctuation is accurate enough to be followed most of the time.