CEFR ONLINE WORKSHOP SERIES 2022 #9

Title

Developing CEFR-based assessment rubrics and rating scales for interaction and production

Date & Time

Thursday, December 1 2022, 16.00 CET

Presenter

Claudia Harsch

Background Reading

To be read *before or after* the workshop.

There are three articles related to the subject of this workshop.

1. Interpretation of the CEFR Companion Volume for developing rating scales in Cuban higher education:
   https://cefrjapan.net/images/PDF/CEFRJournal/CEFRJournal-vol3-5_CHarsch_IPena_etal_Oct2020.pdf

2. Marrying achievement with proficiency – Developing and validating a local CEFR-based writing checklist (see below)

3. Usability of CEFR Companion Volume scales for the development of an analytic rating scale for academic integrated writing assessment (see below)

Please note: the third article is a draft not to be circulated. It will later be replaced with by a link to the published version.

Assessing Writing 43 (2020) 100433

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Marrying achievement with proficiency – Developing and validating a local CEFR-based writing checklist

Claudia Harsch[a,*,1], Sibylle Seyferth[b,1]

[a] University of Bremen, Fachbereich 10, Universitäts-Boulevard 13, 28359 Bremen, Germany
[b] University of Bremen, Germany

ARTICLE INFO

Keyword:
Achievement testing
Constructive alignment
CEFR-based checklist
University context
Proficiency-oriented learning outcomes
Validation

ABSTRACT

Many language course providers face the challenge to align internal, often intuitive assessments to internationally recognised proficiency frameworks for accountability reasons. We report a development and validation project for assessing writing in a university languages centre, where an intuitive, achievement-oriented grading system was aligned to the proficiency levels of the CEFR. We took an iterative approach to developing and validating local writing checklists that combine proficiency-oriented learning outcomes with classroom-based, achievement-oriented assessment goals. Our collaborative approach to checklist development and validation was modelled on the approach reported in Harsch and Martin (2012): Relevant CEFR descriptors were adapted by teachers to fit our context and purpose; the resulting checklist drafts were trialled and revised by teachers in two sessions. Our approach has implications beyond our local context, as it illustrates how such a benchmarking and validation endeavour can be facilitated in a university context where the CEFR has been adapted as the curricular framework. Most importantly, we would like to share our experiences of collaborating in a group of 18 teachers, three course coordinators and two researchers, in order to draw on the expertise of all relevant stakeholder groups, and to ensure that teachers are the central agents in this endeavour.

1. Introduction

Many language course providers face the challenge to move from internal teacher-defined tests and assessment criteria to tests and criteria that are aligned to educational standards or to an internationally recognised framework for accountability reasons. Within such a transition, traditional and intuitive grading systems often have to be reconciled with demands to employ assessment approaches that are standardised to a certain degree, in order to increase transparency and comparability. We report a test revision and development project for assessing writing in a university languages centre, where an intuitive, achievement-oriented grading system was to be aligned to the proficiency levels of the Common European Framework of Reference (CEFR, Council of Europe, 2001). Our particular focus lies on validating our newly developed writing assessment checklists, using a validation approach that combines rater familiarisation, training, trialling and revision, following Harsch and Martin (2012). In line with Barkaoui (2010) and Cohen (1994), we involve teachers as raters in the checklist revision as part of the checklist validation process. As proposed by e.g. McNamara, Hill, and May (2002) and Weigle (2002), we regard sufficient agreement in the application of the checklist as one prerequisite for scoring validity. We take differences in scoring as incentive to discuss underlying reasons, in order to improve the

* Corresponding author.
E-mail addresses: harsch@uni-bremen.de (C. Harsch), seyferth@uni-bremen.de (S. Seyferth).
[1] on behalf of the languages centre team, University of Bremen, Germany

https://doi.org/10.1016/j.asw.2019.100433
Received 25 January 2019; Received in revised form 1 July 2019; Accepted 1 October 2019
Available online 25 November 2019

checklists.

In our context of a large languages centre that serves the four universities in the Land Bremen (Germany), we teach 21 languages. Our language courses are targeting half a CEFR level each, ascending from A1.1 to C1.2 (whereby C1.2 equals C1+). The courses' learning objectives are based on relevant CEFR descriptors. Each course runs over 15 weeks, ending with an informal achievement test. The tests used to be developed by individual teachers; teachers used to score the tests in a variety of ways, some of them using their own intuition-based systems in a mainly deficit-oriented way, i.e., by counting errors and awarding content points. Results of the achievement tests, like all university assessment, have to be reported on the university's 11point numerical scale[2] . Early in 2016, teachers and students demanded that the exams be revised in order to have a greater comparability across different languages, and to transparently report in how far students have achieved the CEFR-based learning goals.

At the languages centre, we formed a collaborative working group that consisted of up to 18 teachers, five course coordinators and two researchers (the authors), in order to revise the existing end-of-course exams. The teachers and coordinators provided insights into teaching and assessment goals, practices and needs. The role of the researchers was to channel the different approaches, insights and practices, to facilitate the establishing of "common" or community practices, and to base teachers' actions in the realm of assessment on research grounds. Henceforth, "we" in this article refers to the collaborative group, as the authors are reporting on behalf of this group.

We were facing the two-fold challenge to develop achievement tests that are transparently aligned to the existing CEFR-based proficiency-oriented learning goals, and to map the university's 11point scale onto the targeted CEFR levels. In this paper, we report how we tackled this challenge for the writing exams, with a particular focus on developing, trialling and validating the assessment checklists. All revisions, the development of the instruments, and the decision-making processes took place in the aforementioned collaborative working group.

We will first discuss relevant literature on developing and validating writing assessment tools, before we briefly describe the developmental work on test specifications and tasks that had foregone the validation endeavours reported in this paper. We then outline the research purpose that is the focus of this paper, i.e. the development and validation of the assessment checklists, which comprised familiarising teachers with the newly developed checklists, trialling the checklists and analysing their applicability, interpretation and feasibility. We present the methodological approach to our validation efforts and discuss the results of the training and trialling sessions with a view to matching achievement testing purposes with proficiency-oriented learning outcomes.

2. Literature review

We will now review theoretical frameworks and projects that are relevant for our endeavour to collaboratively revise our achievement tests; to constructively align learning outcomes, classroom teaching and achievement testing; and specifically to develop and validate a CEFR-based writing assessment tool that is suitable for achievement testing.

2.1. Collaborative approaches to implementing change

Revising an existing exam not only implies developing new tasks and approaches, but it also means implementing changes to an existing system. In order to manage such a change, collaborative approaches in which relevant stakeholders are involved seem most feasible, as they facilitate empowerment, engagement and ensuing implementation. The literature on exam reforms reports several successful projects in which teachers were involved in designing and developing the new exams. There are examples for high-stakes exams, e.g. Holzknecht et al. (2018) for the Austrian Matura revision, as well as for classroom-based assessment, e.g. Studer, Lenz, and Mettler (2004) for a project that targeted formative and summative approaches. The reforms in Germany and Hungary (see Section 2.1 above) also involved teachers in the design and development process. Involving teachers in test development has also positive impact on developing assessment literacy, as e.g. Holzknecht et al. (2018) and Studer et al. (2004) acknowledge, and as is reported e.g. by Baker (2017). While we cannot cover this perspective here, we want to ascertain that the development we report here is part of our broader approach to collaboratively develop assessment literacy in our languages centre (reported in Harsch & Seyferth, in print).

2.2. Constructive alignment

Our endeavour is located within the broader aim of constructive alignment (e.g. Little & Erickson, 2015) of learning goals, teaching content, assessment targets and reporting, whereby the CEFR proficiency system serves as curricular framework. This proficiency system consists of six broad levels of proficiency that describe in positive can-do statements what learners can do and how well they can do it. Since the publication of the CEFR as a descriptive proficiency framework in 2001, many educational systems in Europe have revised their curricula, learning outcomes and assessment aims in order to align them to the CEFR. Exam reforms took place for instance in Austria, where the Matura was revised (Spöttl, Eberharter, Holzknecht, Kremmel, & Zehentner, 2018), in

[2] The numerical scale constitutes an ordinal scale from 1,0 (very good) to 5,0 (fail), with the following non-equidistant steps: 1,0 | 1,3 || 1,7 | 2,0 | 2,3 || 2,7 | 3,0 | 3,3 || 3,7 | 4,0 || 5,0. The lowest mark for a pass is 4,0. These 11 grades are traditionally grouped into the five categories "very good" (1,0 and 1,3), "good" (1,7 to 2,3), "satisfactory" (2,7 to 3,3), "sufficient" (3,7 and 4,0) and "fail" (5,0). This scale is used across all subjects and courses, without any qualitative descriptions.

Germany, where curricula were revised, educational standards introduced and tests developed to regularly monitor educational outcomes (Rupp et al., 2010), or in Hungary, where school-leaving exams were reformed (e.g. Tankó, 2002, for the writing exam). Most reform projects are, however, located in the realm of proficiency testing, be it for high-stakes purposes or educational monitoring; less is reported for adapting the CEFR scales and descriptors for achievement purposes in a specific local classroom setting.

Reconciling achievement testing with proficiency frameworks poses new challenges, for example, when school grades, which serve a multitude of functions, have to be awarded within a context where curricula and learning outcomes are aligned to the CEFR. So far, no convincing solution has been reported in the literature. Some scholars (e.g. Harsch, 2017) argue that school grades should not be mapped onto a proficiency system such as the CEFR, due to the very different purposes of the two systems: The proficiency framework of the CEFR is rather general, abstract, context- and language-independent, and can help judging learners' abilities and development in cases where learners are well-known to their teachers. On the other hand, achievement tests that serve to derive school grades have to be context- and language-specific, targeting concrete achievement objectives. There is little guidance in the literature on how to map local achievement grading systems onto the framework of the CEFR proficiency levels.

Much more is known and reported about aligning proficiency tests to the CEFR (see e.g. COE, 2009; Figueras & Noijons, 2009). With regard to benchmarking writing exams to the CEFR, one can either use the CEFR as starting point; develop test specifications, writing tasks and rating scales that aim to operationalise certain CEFR aspects; and formally link the exam to the CEFR proficiency levels via standard setting methods (see e.g. Harsch & Rupp, 2011 for such an endeavour). Alternatively, existing writing exams can retrospectively be specified and local performances be benchmarked to the CEFR (e.g. Kecker & Eckes, 2010, for the TestDaF exam). Again, most projects reported in the literature refer to high-stakes proficiency exams or educational monitoring contexts; little is known about feasible procedures for reporting local achievement test outcomes with reference to the CEFR proficiency levels.

2.3. Assessing writing achievement

Approaches to writing assessment can broadly be differentiated into multi-level and level-specific approaches, i.e. approaches that span several attainment or proficiency levels vs. those that focus on one specific level (see e.g. Harsch & Martin, 2013 for an in-depth discussion). Multi-level approaches are most suitable when the aim is to gain an overview of the whole range of abilities in a population, or when the focus is on assessing heterogeneous learner groups. Level-specific approaches are feasible when the assessment focus lies on operationalising specific levels of ability or attainment. Hence, for a local achievement test in a classroom setting, where specific learning goals have to be assessed, a level-specific approach seems most appropriate. With regard to task development, particularly in contexts where the CEFR is taken as reference framework, test and task specifications can be derived from relevant CEFR descriptors. For achievement purposes such as ours, the generic CEFR descriptors need to be adapted and complemented by context-specific learning goals.

With regard to rating learner performances, the literature distinguishes holistic, analytic and task-specific approaches (see e.g. Barkaoui, 2011, for an overview). There is no agreement in the literature on which method would yield the most reliable and valid results (e.g. Lumley, 2005). What is known, however, is that analytic approaches allow for detailed feedback, provide more insights into the nature of achievements, counteract potential halo effects and guide raters on focusing on the targeted criteria (e.g. Harsch & Martin, 2013; Smith, 2000). Task-specific approaches, while resource-intensive to develop, have the potential to assess whether learners can master certain task- and genre-related features (e.g. Hamp-Lyons, 1995). Given our context, we set out to explore an analytic and task-specific approach.

Learner performances can be rated by scales or by checklists. Rating scales are used in contexts where it is possible to meaningfully differentiate and describe the targeted attainment levels in one scale, particularly in multi-level contexts. Checklists are more suitable for judging the attainment of specific goals and targets. Brindley (2001), for instance, reports a checklist approach in a classroom setting in Australia, where relevant performance aspects are defined in a level-specific checklist. Teachers then judge learner performances against each statement in the checklist, deciding whether it is *not achieved*, *achieved* or *highly achieved*. Such a checklist approach facilitates achievement-oriented assessment and seems most feasible for our context.

2.4. Developing and validating writing assessment instruments

With regard to the actual development and validation of the checklist that is the focus of this paper, the literature reports of successful involvement of the future users of the instruments in developing and revising them. This is said to lead to more transparent descriptors that can be interpreted more reliably (e.g. Barkaoui, 2010), thus contributing to a valid application and interpretation of the checklist. An iterative validation approach to trialling the checklist while simultaneously training the raters, where raters are also involved in the revisions of the instruments, has been reported as a feasible approach to validation by e.g. East (2009) or Harsch and Martin (2012). With regard to rater training, Knoch (2011), for example, reports that a discussion of rating experiences and timely feedback on rater performance contribute towards efficient training, which in turn enhances rating validity. Such collaborative validation endeavours help tailoring the instruments and approaches to the stakeholders' needs. Furthermore, they lead to empowerment, as the important decisions are taken by those stakeholders who have to implement and use the new system, i.e. in our case our teachers. If teachers who use the instruments are the drivers in the development of these instruments, we assume with e.g. Gallagher & Truly (2008) that such involvement and ownership will lead to a more valid application of the new instruments.

Suitable methods to develop and validate rating scales and checklists encompass intuitive, qualitative and quantitative approaches, whereby performance examples or existing descriptors can be the starting point (see e.g. CoE, 2001, Appendix B for an overview). In our context, where we wanted to revise classroom-based achievement tests across 21 languages, it was not feasible to

describe performance features for 21 languages, particularly not as the aim of the revision was to provide guidance for teachers to achieve a higher transparency and comparability of the exams across the different languages. Moreover, we wanted to align our achievement tests to our CEFR-based proficiency-oriented learning outcomes, so that it seemed most feasible to take existing CEFR descriptors as starting point.

When developing and simultaneously validating new writing assessment tools, not only do aspects of training of the raters, trialling, and revising the instruments have to be taken into account during the development phase, as outlined above. One also has to monitor agreement between raters in the application of the new tool as one prerequisite for scoring validity, as proposed by e.g. McNamara et al. (2002) and Weigle (2002). Some scholars, like for instance Moss (1994), argue for the benefits of acknowledging evaluative diversity stemming from teachers' richly contextualized knowledge of students' learning, such diversity plays a crucial role in formative assessment. In summative contexts, exploring evaluative diversity can contribute to validation endeavours: Validating a new tool aims at establishing a common ground in how the new tool is to be interpreted and applied. This common ground can be monitored by examining the evaluative diversity for its amount of agreement and by exploring underlying reasons for disagreement, in order to improve the tool where necessary.

The development and validation of the instruments is time- and resource-intensive, so that often it is only in high-stakes contexts that new instruments can be fully validated and researched. In classroom-settings, practical constraints may demand certain pragmatic decisions. We will describe in Section 5 below the validation endeavours we could master given our context and constraints.

3. Background

To start with the exam revisions at our languages centre, we took a collaborative approach (for an in depth-description, see Harsch & Seyferth, in revision) in which a group of teachers, course coordinators and researchers developed test specifications and writing tasks that operationalise the CEFR half-levels targeted by our courses (see Introduction above). Since we teach 21 languages at the languages centre, we aimed at test specifications that could serve as an orientation frame across all languages, while offering spaces for individual adaptations. This space for adaptions acknowledges the experiences and the rich contextual knowledge of our teachers (e.g. Gallagher & Turley, 2012), which was also to be reflected in the assessment. The test specifications development and the challenges we faced when adapting the CEFR to a university context across different languages are reported in Seyferth, Lavagno, Kucera, and Harsch (2018). In addition to the test specifications across all languages taught at our centre, we needed to account for the special status of one language at our language centre: English is the language where most courses are taught, with the largest group of teachers. Our English courses target academic purposes (EAP) at CEFR levels B2.1 to C1.2. Hence, we decided to adapt and translate the initial test specifications for our English courses. For the translation, we used the original CEFR descriptors in English and the learning objectives from the *Global Scale of English for Academic Purposes* (Pearson, 2018).

Once we had the specifications in place, we needed to map the university's 11point scale onto the targeted CEFR levels, in order to match the CEFR-based proficiency-oriented learning outcomes with the university's achievement grades. This adds transparency and meaning to the otherwise meaningless 11point university achievement scale. Clarifying what constitutes a "pass" and when a" full mark" could be awarded would also help to decide whether a multi-level or level-specific approach to rating would be more feasible, and whether a rating scale or a checklist approach would be best. In July 2016, a group of 16 teachers, 2 coordinators and 2 researchers conducted an initial benchmarking session, in which we mapped the existing 11point numerical university scale onto the CEFR-based learning outcomes. We concentrated on one CEFR half-level, a course located at B2.1, in a language common to all our teachers. We analysed three student texts that were previously elicited by an existing B2.1 task (a task was chosen that aligned well with the newly developed test specifications). These student texts had been graded previously, using the university scale. This approach was chosen as we did not have time to implement the new tasks, and because we wanted to ensure that the transition to the new approach would be as smooth as possible. For our analysis, we used the B1+, B2 and B2+ descriptors of selected CEFR writing scales (CoE, 2001, pp. 61, pp.82), the scales for linguistic competences (CoE, 2001, Chapter 5), and the assessment grid from the *Manual* (CoE, 2009, pp. 187). The aim was to determine what aspects had to be met to award a "pass" (as expressed by a 4,0 on the university scale), and what aspects had to be met in order to achieve "full marks" (a 1,0 on the scale).

The outcome was unanimous: All participants agreed that in order to achieve full marks, all relevant aspects of the targeted level had to be met; all participants argued that it was not necessary to show any of the required features of the next half-level. With regard to the pass mark, all participants agreed that while it was acceptable that the student texts showed some aspects of the level below the targeted level, the texts should show more of the features stated at the targeted level than of the level below. In other words, a text located at the beginning of the targeted level constitutes a pass (grade 4,0), a text located at the top of the targeted level constitutes a full mark (1,0). Fig. 1 depicts the mapping results:

The arrow indicates the spread of the targeted level; the bars indicate features located at the top end respectively at the beginning of a level. The fact that we have to use 10 grades of the university scale (from 1,0–4,0; the *fail* grade was unanimously located below the targeted level) within one half-level of the CEFR implies certain challenges. From a practical perspective, for example, it was not possible for us to develop such fine-grained descriptors that could mirror ten grades within one CEFR half-level. Hence, it seemed more manageable to describe the target, i.e., translate the proficiency-oriented learning outcomes into assessment-oriented checklist descriptors. Such analytic, qualitative checklists help expressing our shared values and expectations, both with regard to guiding the assessment and giving qualitative feedback to our students (see e.g. Broad, 1994).

We then needed to find a way to match the qualitative checklists with the quantitative university grading system. In order to achieve this, we decided against a yes/no checklist approach; rather, we decided to rate the checklists statements against the four major grading categories (see FN 1 above). Furthermore, we needed spaces for adding relevant aspects from the classroom, to cater

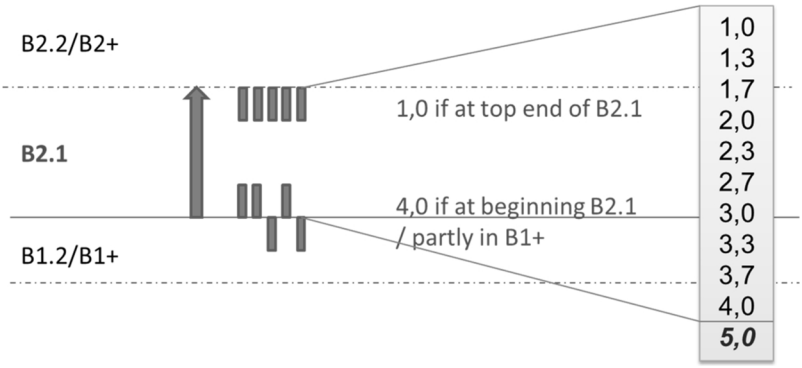

Fig. 1. Mapping the university scale onto the CEFR.

for the achievement purpose of the exam and for task-specific features. Having laid the groundwork by developing test specifications and by mapping the university achievement scale onto the targeted CEFR half-levels, we now turn to the main purpose of our paper, i.e., developing and validating a suitable checklist for our writing achievement tests.

4. Aims and methodology

4.1. Aims

Based on the collaboratively developed multi-language test specifications and writing tasks, and based on the initial mapping of the university grading system onto the targeted CEFR levels, we explored how a proficiency-based checklist for achievement purposes could be developed and validated, that

- is adaptable to fit individual teachers' needs and classroom practices,
- reflects task-specific aspects,
- is practicable for classroom assessment and reporting,
- can be reliably and validly used by teachers,
- allows transparent alignment of achievement testing in a classroom setting with CEFR-based proficiency-oriented learning outcomes,
- and allows reporting on the university's 11point numerical scale.

4.2. Methodology

Our validation endeavours aimed at gaining insights into the feasibility of the new approach, the usability of the new checklists, the consistency with which the checklists can be applied and interpreted, and the comparability of the resulting scores; we also aimed at monitoring the comparability of the old and the new approaches. We took a collaborative approach to checklist development and validation in three steps, starting with drafting the checklists, based on our tests specifications and relevant CEFR descriptors. The checklists were then refined in two trialling / training sessions, following the validation procedures reported in Harsch and Martin (2012). The stakeholder groups of teachers, course coordinators and researchers were involved in all steps, whereby the teachers were the drivers in the process, constituting the largest group with the biggest impact on decisions. Fig. 2 outlines the process:

Step 1 constituted of drafting the checklists. In step 2, i.e. the first of the two one-day trial / training sessions in January 2018, we focused on the common language of the country in which the university is based; the outcomes led to further refinements of the checklists. In step 3, i.e. the second session in February 2018, we formed smaller groups around the other most commonly taught languages at our centre, i.e. English, Roman languages (French, Italian, Portuguese), and Slavic languages (Polish, Russian). In each of the two sessions, teachers[3] rated three selected student texts (elicited by new tasks), using the new checklists. We restricted the selection to three student texts due to time constraints, and because we wanted to use (dis)agreement as stipulation to locate and discuss underlying reasons for variation. During the discussions, we not only formed a shared understanding of what constitutes a "good piece of writing", but we also formed shared expectations with regard to the targeted CEFR levels. In addition, we located ambiguous formulations in the checklist wording so that the discussions also led to subsequent revisions of the checklists.

In step 3, the teachers also scored the three student texts in the traditional way, in order to examine potential variations between the traditional and the new approaches, and to ensure a certain amount of consistency between the two assessment approaches. Again, we took differences in our assessments as starting point to discuss underlying reasons. These discussions refined and furthered

[3] Teachers developed the test specifications and the tasks; they trialled the tasks in their classrooms and selected the student texts for the benchmarking sessions.

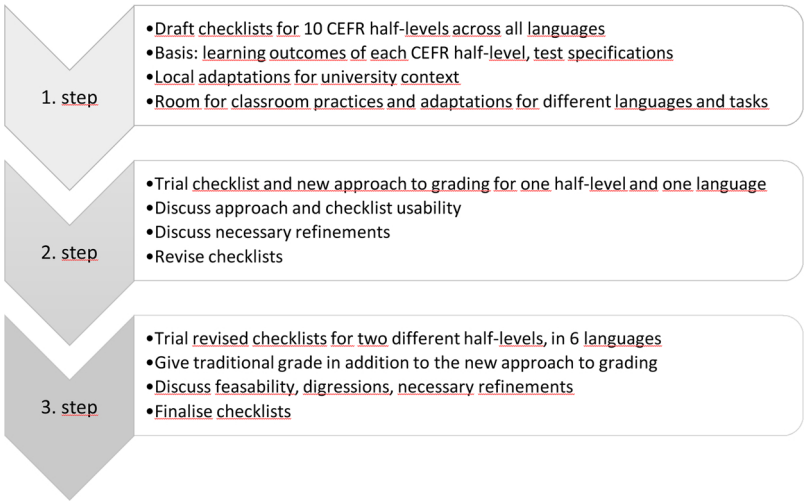

Fig. 2. Development, trial and refinement of CEFR-based achievement checklists.

our shared understanding of expectations, CEFR levels and checklist interpretation, and they facilitated refining the checklists further. During the discussions, there was sufficient time to voice concerns regarding the transition from intuitive scoring, which often focused on error counting, to rating student texts against proficiency-based descriptors in a checklist.

4.3. Limitations

With regard to analysing the rating data that we collected during steps 2 and 3, the small scale of our data set (three student texts per session and language group) precludes inferential statistics or applying IRT models (see Wilson, 2008). Hence, we have to confine our analysis to descriptive methods, which nevertheless suit our purpose well, i.e. identifying (dis)agreement in order to stipulate discussion of underlying reasons. Monitoring applicability and consistency with a larger set of data will be conducted once the checklists are applied with larger groups of students. The purpose of the research reported here is to ensure that the checklists are ready to be implemented in the classroom. This may seem as a limitation, but given the constraints classroom-based assessment projects usually experience, we would like to share a feasible approach for practitioners who do not have the resources of the international testing industry.

5. Validation procedures: development and trialling of CEFR-based achievement checklists

We now describe each of the three steps in turn, presenting and discussing outcomes of the first step (see Fig. 1 above), while we present aims, methods, data and findings for the second and third step.

5.1. Step 1: drafting the checklists

During 2016/17, the group of teachers, coordinators and researchers collaboratively developed assessment checklists that operationalised the targeted CEFR half-levels. Since the exams are achievement tests, we needed to ensure that local classroom practices could be aligned with the CEFR-based learning goals. In addition, since we teach 21 languages at the languages centre, we needed checklists that could be applied across all languages, while offering adaptation options to cater for requirements of the different languages. For instance, teachers of languages with a non-Latin script required that closer attention be paid to orthography on the lower levels. We also wanted spaces for individual adaptations to cater for individual classroom practices.

The development spread over the course of a whole year because it took place parallel to the development of new writing tasks, and parallel to the normal teaching. While teachers received a reduction of their teaching load, it was not possible to take time out in order to develop all checklists at once. We started with the lowest level A1.1 and worked our way up to C1.2. The language of the checklists was the language of the country in which the languages centre is based, as this is the common language shared by all staff members.

The basis for the checklist development were the achievement test specifications, in which learning objectives, task characteristics, expected learner text features and relevant assessment criteria had been defined previously by the collaborative working group. In addition to the test specifications, we took all relevant CEFR scales into consideration. Based on discussions within the group and driven by our teachers' knowledge and experiences, we differentiated the following assessment criteria: *task fulfilment* (differentiating topic/length and genre/style), *organisation*, *vocabulary*, *grammar*, *orthography*. For each criterion, a number of descriptors were formulated, in a terminology that we tried to keep positive and as simple as possible. We left spaces so that teachers could briefly specify the tasks, the text types, the topic-related vocabulary demands and the grammatical aspects that they were expecting, based on their

Task dimension		Completely fulfilled	Almost fulfilled	Largely fulfilled	Partly fulfilled	Not fulfilled
Task fulfilment	Can complete the **task** in terms of content and writes 80-100 words **about:** *[specify topic]*					
	Genre *[specify requirements]*					
Linguistic dimension		Very good ++	Good +	Satisfac-tory +/-	Sufficient -	Insufficient --
Organisation	Can write a series of simple phrases and sentences and link them with simple connectors such as *and, but, or, because, when, if*.					

Fig. 3. Initial checklist draft, level A2.1.

teaching. Since the initial checklists were drafted in the language of the country in which the university is based, we present a rough translation of the draft for level A2.1 in Fig. 3, illustrating the criteria *task fulfilment* and *organisation* (the spaces for individualization and specification are indicated in *[Italics]*):

In terms of the assessment procedure, each student text is to be analysed against each of the criteria, whereby each criterion, as defined by the descriptors, is to be rated on a 5point scale, in order to match the five major grading categories of the university system (see FN1 above). The five columns in Fig. 3 represent these five categories. The idea is that teachers tick the respective column for each criterion in an analytic approach, and then form their overall grade based on the analysis of the categories. This overall grade is expressed as one of the 11 grades of the university grading system. The checklist is to be filled in by teachers and handed out to students. This way, teachers and students can transparently see in how far the learning objectives have been achieved, and they can visually understand how the overall grade was derived from the analysis of the student text against the learning objectives.

5.2. Step 2: first trial and training session

5.2.1. Aims and method

In the first trial and training session in January 2018, which lasted two hours, the aim was three-fold: First, we wanted to familiarise as many teachers as possible with the new checklists and the rating approach, because not all teachers could or wanted to participate in the developmental work during step 1. Second, we wanted to trial the checklists, the analytic rating procedures and the new approach to derive the achievement grades, in order to examine whether teachers found this approach feasible. Here, we monitored consistency and discussed differences within the group. The third aim was to gain feedback on the checklist wording and conceptualisations, in order to revise where necessary. The procedure of familiarisation, trial and revision contributes towards validating the new checklists for their achievement testing purpose (see Harsch & Martin, 2012).

For practical reasons, we focused on the common language of the country in which the university is based, and on a level that the majority of our teachers were familiar with, i.e. A2.1. A new writing task operationalising the test specs had previously been developed by the group. It was trialled in an A2.1 course, in order to collect student texts. Three student texts were pre-selected collaboratively by a teacher and a researcher. The texts represented a strong, medium and weak performance. 13 teachers and 2 coordinators participated in this session that was facilitated by the researchers (i.e., the authors).

First, participants were asked to familiarise themselves with the checklists, and we explained the rating approach. We then introduced the task and the three student texts. Participants were asked to analyse the student texts in terms of their achievement of the criteria; participants were to tick the respective column next to the criterion under analysis. Once they had analysed all criteria, they were to form their overall judgment; we deliberately did not give guidance for weighting the criteria, as we wanted to explore how participants would intuitively weigh them. After a break, we presented the results and discussed participants' experiences, perceptions, underlying reasons behind differences and necessary refinements to the checklist wording.

5.2.2. Results and discussion

Table 1 shows the aggregated rating results for the three student texts for the analytic criteria. The ratings were coded 1 to 5 for each of the five columns, from 1 – *very strong / completely fulfilled* to 5 – *insufficient / not fulfilled.*

The results show that there is a substantial amount of variation in all criteria. To examine the sources of these variations, and to examine variation in the overall grades, we now present rating and grading results for Text B, as this was the text with the largest variations. Table 2 shows the results for this text, broken down for the 15 participants.

The results show that different raters differed in their judgements in different criteria. For the linguistic criteria, raters showed a range of four rating categories, which constitutes a rather large variability. Moreover, some raters did not use one of the four categories but rather placed their tick between adjacent categories; these raters reported that they felt unable to decide for one or the other. Raters 10 and 11 did not provide an overall grade, as they found the transition from rating to giving a grade too challenging at that time.

Table 1
Rating results Session 1 across all texts.

	Text A - medium			Text B - weak			Text C - strong		
Criteria*	mean	mode	SD	mean	mode	SD	mean	mode	SD
TL	1.33	1.00	0.59	3.17	3.00	0.82	1.43	1.00	0.56
GS	1.70	1.00	0.84	4.80	5.00	0.41	1.40	1.00	0.51
ORG	1.90	2.00	0.60	3.13	3.00	0.74	1.23	1.00	0.42
VOC	1.70	2.00	0.59	2.80	2.00	0.90	1.13	1.00	0.35
GRA	2.60	3.00	0.74	3.07	4.00	0.92	1.87	2.00	0.52
ORTH	1.67	1.00	0.72	2.20	2.00	0.86	1.27	1.00	0.46

* TL: topic/length; GS: genre/style; ORG: organisation; VOC: vocabulary; GRA: grammar; ORTH: orthography.

Table 2
Rating results Session 1 Text B.

Criteria	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15
TL	3	**1.5**	3	3	3	3.5	**1.5**	3	4	4	3	4	4	3	4
GS	5	5	4	4	5	5	5	5	5	5	5	5	4	5	5
ORG	3	3	3	2	5	3	3	2	3	4	3	3	3	4	3
VOC	2	3.5	4	2	4	3.5	3.5	3	3.5	3	3	2	2	2	**1**
GRA	3.5	2	3	**1**	4	2.5	3.5	4	4	4	2.5	4	3	3	2
ORTH	1	2	3	1	2	2	3	3	2	**4**	2	3	2	2	1
Criteria mean	*2.92*	*2.83*	*3.33*	*2.17*	*3.83*	*3.25*	*3.25*	*3.33*	*3.58*	*4.00*	*3.08*	*3.50*	*3.00*	*3.17*	*2.67*
Overall Grade	3,3	2,3	3,0	**1,7**	3,3	3,3	3,3	3,7	3,7	n.g.	n.g.	**4,0**	3,3	3,7	3,3

Interestingly, for those 13 cases where a grade was given, the arithmetic means of the criteria ratings (which we calculated after the grades were given) show a rather close relation to the grades that the participants gave; this could be interpreted as an indication of internal consistency. We had explained beforehand to the participants that they should not start calculating the mean, in order to prevent them from going back to allocating points, as was the approach used in the past. Rather, participants were encouraged to reflect on the different profiles, strengths and weaknesses emerging from the criteria-based analysis, and then come to a balanced overall grade based on their rich experience and knowledge; thereby, they should use the emerging profile in the four grading categories as guide to deriving the overall grade. In this session, we had not yet decided on the weighting of the different criteria, and we wanted to explore how participants intuitively weighed them. For the case of Text C presented in Table 2 above, the discussions revealed that the non-fulfilment of the required genre swayed many teachers towards a lower grade than the arithmetic mean would have suggested. In this session, we did not yet come to a final decision on how to weigh the criteria.

When we discussed the rating results in the group, we found two main reasons for the differences in judgements:

- different understanding and interpretation of the descriptors due to wording, conceptualisation or lack of coherence in the checklist;
- different approaches to interpreting and weighting errors, with some raters sanctioning every error, and others taking the approach outlined in the checklist, where certain kinds of errors are acceptable for a certain level of proficiency.

The first issue could be addressed by refining the checklists, as outlined in section 5.2.3 below. The second issue needed to be addressed not only by clarifying in the checklists how errors should be treated, but also by more in-depths discussions of what constitutes a positive approach[4] to assessment. This included discussing how learner language is conceptualised positively in the CEFR, i.e. as a valid language variety that can be described by what learners already can do. This conceptualisation seemed to be in contrast to the prevailing conceptualisation with some participants, i.e. that learner language is characterised by errors and missing features. These discussions were deepened between the two trial sessions, when we revised the checklists; they also came up in the second trial and training session, as well as in several meetings over the ensuing months. These discussions seminally enhance assessment validity, as they contribute to forming a shared understanding of what constitutes "good" writing at different CEFR levels for our context, as well as to understanding the positive approach of rating in contrast to allocating points or counting errors.

[4] By positive approach, we are referring to an approach that focuses on what learners already can do and how well they can do it, rather than focusing on errors, deficits and what is missing. In our context, many teachers traditionally focus on errors, on what is missing, on deducting points – hence, we needed to reorient ourselves to focus on what learners (positively) already can do. This is also the approach that the CEFR takes, as illustrated by its can-do statements that describe in a positive way what learners can do. We are trying to apply this approach rather than marking all errors with a red pen, as was done traditionally.

Task dimension		Completely fulfilled	Almost fulfilled	Largely fulfilled	Partly fulfilled	Not fulfilled
Task fulfilment	Can complete the **task (e.g. email or blog)** in terms of **content** and writes 80-100 words **about:** *- [specify topic]*					
	Genre The **communicate goal** becomes clear: *- [specify goal]* *If required: greeting, opening and salutation are appropriate.*					
Linguistic dimension		Very good ++	Good +	Satisfac-tory +/-	Sufficient -	Insuffi-cient --
Organisation	Writes a series of **simple** phrases and sentences. Can link simple sentences and uses mean of cohesion (references within the text) and/or **frequent** connectors such as *and, but, or, because, etc.*					

Fig. 4. Checklist draft 2, level A2.1.

5.2.3. Revision of checklists

Based on the discussions of the first session, the group collaboratively revised the checklists, aiming at creating more clarity and coherence across all levels. The conceptualisation of the criterion *task fulfilment* was revised, adding communicative goals. The wording of some descriptors was changed to clarify their meaning; descriptors that contained several ideas were split into separate sentences. Due to the second issue mentioned above in section 5.2.2, we added statements for the language criteria that clarified which kinds of errors were considered acceptable, based on group discussions. For instance, at level B2.1 for the criterion *vocabulary*, we added the statement: "Some errors in vocabulary choice and usage are allowed as long as they do not hinder communication."

Fig. 4 shows the revisions for level A2.1, illustrating the criteria *task fulfilment* and *organisation*, again presented here in a rough translation:

Once the checklists had been revised, a smaller group of English teachers and coordinators adapted them for EAP, following the approach chosen for the test specification (see section 2 above). For this adaptation and translation, we used the English test specifications, the original CEFR descriptors in English, as well as the GSE descriptors for EAP (Pearson, 2018). Appendix A shows the initial checklist draft for English, level B2.1, which we used in the second trial and training session.

5.3. Step 3: second trial and training session

5.3.1. Aims and method

The second session in February 2018, which lasted a whole day, aimed at extending the checklist trial and validation to include more levels and more languages. This served to deepen the initial familiarisation, and to gain further feedback on the usability and validity of the checklists in order to finalise them. Moreover, we wanted to compare the new approach with the traditional grading, in order to facilitate a smooth transition into the new system, and to encourage teachers and coordinators to reflect on differences between the former, mostly intuitive counting approach and the new rating approach that focussed explicitly on our curricular, CEFR-based learning outcomes. The training took place at the end of the semester, to prepare teachers for grading the end-of-term exams with the new approach.

A total of 15 teachers and 2 coordinators participated, 10 of whom had also participated in the first trial/training session. We formed smaller groups around the most commonly taught languages, i.e., English (n = 9 raters), Roman languages (French, Italian, Portuguese; n = 5 raters), and Slavic languages (Polish, Russian; n = 3 raters). For English, a new task had been developed by one of our teachers, who had also trialled it and had selected three student texts. For the other languages, teachers had collaboratively developed a writing task for level A2.1, which they then had translated into their respective languages and trialled in their courses. Teachers had selected three student texts per language from these trials.

In analogy to the first session, participants were asked to familiarise themselves with the revised checklists, and we explained the approach once more. Participants then worked in their respective groups, analysing the selected student texts as explained above in section 5.2.1. Again, we left the final weighting to derive the overall grade to the participants. Finally, in addition to coming to a grade based on the new approach to rating, participants were asked to give a grade in the way they used to do traditionally. After a break, the facilitators presented the results and we discussed the experience, perceptions, reasons for differences and necessary refinements in the group.

Table 3
Rating results Session 2 across three texts for English.

	Text A - medium			Text B - weak			Text C - strong			
Criteria*	mean	mode	SD	mean	mode	SD	mean	mode	SD	*SD rev.*
TF	2.72	3.00	0.97	4.78	5.00	0.44	1.50	1.00	1.00	*0.37*
GS	2.81	3.00	0.53	4.63	5.00	0.52	1.43	1.00	0.79	*0.41*
ORG	3.00	3.00	0.90	4.83	5.00	0.35	1.33	1.00	0.66	*0.23*
VOC	2.61	3.00	0.60	4.83	5.00	0.35	1.22	1.00	0.67	*0.00*
GRA	2.78	2.00	0.87	4.69	5.00	0.37	1.39	1.00	0.78	*0.53*
ORTH	2.00	2.00	1.00	3.88	4.00	0.83	1.44	1.00	0.73	*0.74*

* TF: task fulfilment; GS: genre/style; ORG: organisation; VOC: vocabulary; GRA: grammar; ORTH: orthography.

5.3.2. Results and discussion

Here, the results are illustrated for the group of 9 participants who worked with the English task at level B2.1, as the other groups were very small in numbers. Table 3 shows the aggregated rating results for the three student texts for the analytic criteria. The ratings were again coded 1 to 5 for each of the five columns, from 1 – *very strong / completely fulfilled* to 5 – *insufficient / not fulfilled*.

The results show a rather high level of heterogeneity, particularly for Text A. For Text C, we had one "outlier"[5] ; once this outlier was removed from the data set, the standard deviation was greatly reduced, as shown in the last column in Table 3. In order to examine potential reasons for variations, and to examine variations between the new and the traditional grades, we present the individual rating and grading results for Text A, as this was the text where we found the largest variations. Table 4 shows the results for this text, broken down for the nine participants.

With regard to differences between raters, there are relatively few and unsystematic instances where individual raters differed from the group. In the discussions, these differences could be explained and in most cases, the "differing" raters could understand and follow the arguments of the other raters in the group. In this phase, differences had their reasons mainly in how participants treated and interpreted errors.

When it comes to comparing the grades based on the checklist analysis with the grades derived traditionally, the two grades for Text A are always adjacent, in two cases even the same. There is no systematic trend toward more lenient or harsher grading with the new approach. Encouragingly, the grades derived with the new approach seem to be rather consistent with the arithmetic mean, apart from one case (R20). Looking at the other two texts, the grades are quite similar, but with a slight tendency towards the new grades being one point better on the university scale.

The results for the other two language groups were similar, with a few differences that group discussions could solve or explain. The similarities between the new and the traditional grading systems showed here, as well, with a tendency of the new approach to yield slightly better grades (the majority of grades were the same, with some of the new grades being one point better in the university grading system).

The results were presented to the three groups; the first round of discussions took place within the language groups. The groups then reported their discussion outcomes in the plenum. Across all groups, teachers and coordinators perceived the new approach as feasible, adding transparency to the grades and allowing for detailed feedback, as it is intended to hand back the checklists to the students. The slight trend towards better grades was explained by those teachers' focus on what students can do rather than focusing on errors. The discussions revealed that teachers regarded the better grades for the texts in question as justified, given the learning objectives at hand. Participants felt prepared to try out the new checklist approach when marking the end-of-term exams. Nevertheless, there were some critical voices who were taken seriously, and who could be convinced by their colleagues to at least give the new approach a try.

The following issues emerged that needed to be taken care of in a final round of revisions:

- the criterion *task fulfilment* was perceived as not sufficiently reflecting the complexity of the tasks and requirements;
- the statements referring to the treatment of errors were perceived as not yet clear enough;
- particularly in the language criteria, participants criticised that the "can do" statements were misleading, as the assessment focus lies on one text sample rather than on students' overall abilities;
- some participants reported that particularly for the criterion *organisation*, they focused on the examples given for connectors and tended to ignore other cohesive devices;
- participants asked for more guidance on how to weigh the two main aspects, i.e., task dimension and language dimension;
- the format and layout was criticised as not reader-friendly enough; participants requested a comment field where they could add individual feedback to the students.

[5] In the discussions, all participants had space to voice their reasons and concerns. Nobody was forced to agree with the group; rather, the focus was on finding out what drove colleagues to give a certain rating, and to establishing a shared understanding of what constitutes a certain rating.

Table 4
Rating results Session 2 Text A English.

Criteria	R3	R4	R12	R13	R16	R17	R18	R19	R20
TF	2	3	**4.5**	3	3	2	3	3	**1**
GS	3	3	3.5	3	–	3	2	3	2
ORG	2	3	3	3	3.5	2	3	2.5	**5**
VOC	**1.5**	3	3	3	3	2	2	3	3
GRA	2	2.5	2	3.5	2	3	2	4	4
ORTH	2	1	1	2	1	2	2	**3**	**4**
mean	*2.08*	*2.58*	*2.83*	*2.91*	*2.50*	*2.33*	*2.33*	*3.08*	*3.17*
Grade / checklist	2,3	2,7	3,0	3,0	–	2,3	2,3	2,7	2,3
Grade traditional	2,7	2,3	3,0	3,3	3,0	2,3	2,0	2,3	2,7

Task fulfilment (when in doubt, this carries more weight)		Completely fulfilled	Almost fulfilled	Largely fulfilled	Partly fulfilled	Not fulfilled
Content	Can complete the task in terms of content and writes about: *- List content expectations.*					
Language functions, comm. effect	Can realize the following language functions: *- List targeted language functions here (e.g., introduce oneself, invite, describe).* Achieves the intended communicative effect: *- Briefly specify the communicative goals (e.g. describe an experience so that it becomes understandable).*					
Genre	Shows the required genre features: *- List genre requirements (e.g. E-Mail or blog … e.g. greeting, opening, closing).*					
Language Competences		Very good ++	Good +	Satisfac-tory +/-	Suffi-cient -	Insuffi-cient --
Organisa-tion	Writes a series of simple phrases and sentences. Can link simple sentences and uses mean of cohesion (references within the text) and/or frequent connectors.					

Fig. 5. Checklist final version, level A2.1.

5.3.3. Final revisions

Based on the outcomes of the second trial and validation session, a smaller group of teachers, coordinators, and researchers revised the checklists once more. The conceptualisation of *task fulfilment* was revised completely, now differentiating the three criteria 1) content; 2) language functions / communicative effect; 3) genre / register. The reference to text length was dropped as teachers agreed that it need not be controlled: texts that are too long or too short may get lower achievement in the criterion *communicative effect*. In the language criteria, statements referring to errors were put in Italics, to indicate acceptable errors that are not to be sanctioned. In the language criteria, the "can do" phrases were reworded to better describe text features rather than focus on student abilities. In *organisation*, the examples for connectors were dropped as they led teachers to search for these, ignoring other coherence features. We also added statements on the weighting of *task fulfilment* vs. *language competences*, whereby the group decided that up to level B1+, *task fulfilment* would carry more weight if in doubt, while from B2 onwards, *language competences* would bear more weight if in doubt. Furthermore, we designed a template in Word with text fields where teachers can insert specifications to adapt the template for specific tasks and for the learning objectives to be achieved in a specific course. We also added a comment space in the template. Finally, the checklists were formatted in a reader-friendly layout. Fig. 5 illustrates the changes for the criteria *task fulfilment* and *organisation* at level A2.1 (again in a rough translation).

The grey fields in Fig. 5 indicate the spaces in the template where teachers can add their specifications. An example for a complete checklist is given in Appendix B, where the results of the revisions are illustrated for English level B2.1.

6. Conclusions

Our approach to collaborate in a group of teachers, coordinators and researchers on the development and validation of new checklists for assessing writing allowed us to develop tools that validly reflect learning outcomes and classroom practices, and that

can be applied and interpreted with sufficient agreement and consistency. Adapting relevant CEFR descriptors for achievement-oriented assessment checklists facilitated the constructive alignment of proficiency-oriented learning goals with achievement-oriented assessment targets. Furthermore, by providing spaces for the teachers to specify task demands and teaching contents, the checklists also align the assessment to the realities of the classroom. Thereby, we managed to "marry" achievement tests with proficiency-oriented learning goals. The first step was to map the local grading system onto the targeted CEFR levels. Next, due to the characteristics of the local system, we chose a checklist approach in which we described the assessment targets with close reference to the targeted CEFR-based learning goals. Judging how well students have achieved the specific goals with reference to the university's ordinal grading system adds qualitative proficiency-oriented transparency to the numerical reporting of the achievement tests. Our approach focusses on what students already can do and how well they can do this, with close reference to what was taught in the courses. The checklists provide a common frame that leaves space for teachers to account for differences in different language systems and in teaching content and approaches.

The collaborative development of test specifications, tasks and checklists, in which teachers were the drivers and took the important decisions, facilitated the transition to a new assessment system. Introducing the CEFR-based checklists for achievement purposes in a combined trial and training approach facilitated familiarisation with the new system and the checklists; allowed targeted discussions and feedback; showed where the checklists needed refining; and most importantly helped teachers voice their concerns regarding the transition from intuitive grading to rating with a checklist. This trial / training / revision approach also served as initial validation of the checklists, to ensure that teachers can form a shared understanding of the expected writing achievements, and apply and interpret the checklists in a comparable way across languages and courses. While including important stakeholders in the development and validation process may be time-consuming and resource-intensive, it is an effective way to establish acceptance and empower teachers to develop their own tools. Moreover, it enhances scoring validity, contributes towards developing assessment literacy in a local community of practice (Harsch & Seyferth, in revision), and enhances acceptability among stakeholders such as teachers and students (Harsch & Seyferth, in print).

The insights we have gained have implications beyond our local context. Many language course providers face the challenge to move from intuitive criteria that are not explicitly defined or described, to comparable and transparent criteria that are aligned to an internationally recognised framework. In such contexts, our approach can illustrate how such a transition can be managed, and how the expertise of teachers, course organisers and researchers can be drawn upon in a synergetic endeavour to develop and validate the new tools. Moreover, our approach to developing a proficiency-based checklist adaptable to fit individual teachers' needs and classroom practices can serve as example for contexts where achievement testing is to be aligned to a proficiency framework.

We have to concede that our validation efforts were constrained to the development phase and to very small samples. Nevertheless, they allowed us implementing a tool that achieved satisfactory acceptance among the teachers and showed acceptable comparability. Future research needs to be carried out with larger data sets, as suggested by one of the anonymous reviewers, to examine rating consistency via multi-faceted IRT models, and to explore the dimensional nature of the criteria, in order to investigate how different dimensions of the checklists work and how teachers interact with those dimensions. Examining the quality and usefulness of the feedback generated by teachers' or students' use of the checklist is a further avenue of future research, which we are currently planning. Another promising line to pursue is to explore how teachers evaluate texts composed under authentic conditions, as proposed by another reviewer. Here, a developmental, process-oriented angle could be taken, in which students choose their rhetorical situations (topic, genre, audience, purpose, media, forum, etc.) and develop their performances over time with research, revision, and response as part of their composing processes. Surveying teachers and students about their perceptions of summative and formative writing assessment could yield powerful insights into what constitutes effective assessment in authentic settings.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

None.

Acknowledgements

We would like to thank all teachers and coordinators at the languages centre who participated in this project and in the benchmarking and training sessions. We appreciate the dedication of our languages centre team towards implementing a new approach to assessment, and their trust in managing this process of change together. We would also like to thank two anonymous reviewers for their insightful comments and suggestions that helped improving the manuscript. All remaining errors are ours.

Appendix A

See Table A1

Table A1
Initial draft of Checklist for EAP, level B2.1, after first trial / training session.

Task fulfilment		Completely fulfilled	Almost fulfilled	Largely fulfilled	Partly fulfilled	Not fulfilled
Task fulfilment	Can complete the **task** in terms of content and communicative purpose and write **250 - 350 words.** *[Specify task: descriptions of processes and graphs; reports; formal letters, formal emails; job applications (motivation letters, CV); simple academic writing (discursive)]* Output text demonstrates candidate's ability to: *[Specify targeted language functions, e.g., describe, report, contrast, compare, justify, evaluate, argue]* **Genre, format, style** Can use correct academic style, register, format and conventions appropriate to genre and task *[Specify expected task-specific characteristics here.]* Demonstrates ability to incorporate culturally appropriate conventions. *[Specify expected conventions here.]*					

Language competence		Very good ++	Good +	Satisfac-tory +/-	Sufficient –	Insufficie-nt –
Organisation (Coherence, Cohesion)	Organises text using **coherent and cohesive paragraphs** using topic sentences and supporting details and relevant examples. Develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. **Text structure** is clearly marked using a limited range of **discourse markers** to signal rhetorical functions: *[Specify. e.g. cause and effect relationships, difference between fact and opinion, similarities and contrasts between two ideas, problem and solution relationships]* Organises text according to **appropriate formats and sections,** where necessary with headings.					
Vocabulary (Range, usage)	Demonstrates **sufficient command** of vocabulary required for the task. *[Specify target vocabulary relevant to task here.]* Varies formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution. Vocabulary usage is **generally accurate**. Some errors in vocabulary choice and usage are allowed as long as they do not hinder communication.					
Grammar (Range, Accuracy)	Demonstrates a **good command of grammar** and the ability to use a variety of sentence structures. *[Specify target grammar relevant to task here.]* Non-systematic errors and mistakes in sentence structure may still occur.					
Ortho-graphy	Spelling and punctuation are **reasonably accurate**. Occasional errors **may still appear**, especially for punctuation, and resulting from L1 interference.					

Appendix B

See Table B1

Table B1
Final Checklist for EAP, level B2.1, after second trial / training session.

Task fulfilment (language criteria carry more weight if in doubt)		Completely fulfilled	Almost fulfilled	Largely fulfilled	Partly fulfilled	Not fulfilled
Content	Can complete the task in terms of content: *- List expected content point here.*					
Language functions, comm. effect	Output text demonstrates candidate's ability to realize the following language functions: *- List target language functions here (e.g. describe report, contrast, compare, justify, evaluate, argue).* Achieves the targeted communicative purpose and effect: *- List expected effects here.*					
Genre, register	Can use appropriate format and conventions for genre and task: *- List expected genre characteristics here.* Chooses appropriate register: *- Specify formal / informal / neutral.* Demonstrates ability to incorporate culturally appropriate conventions: *- List expected conventions here.*					

Language competence (l criteria carry more weight if in doubt)		Very good ++	Good +	Satisfactory +/-	Sufficient -	Insufficient –
Organisation (Coherence and Cohesion)	Organises well-structured, coherent and cohesive text (e.g. using topic sentences and supporting details and relevant examples). Text is organized in paragraphs. Text structure is clearly marked using a limited range of discourse markers to signal rhetorical functions. *Longer texts may still occasionally be "jumpy".*					
Vocabulary (Range and Usage)	Demonstrates a relatively large vocabulary range required for the task: *- Specify topic of task here.* Varies formulation to avoid frequent repetition. *Lexical gaps can still cause hesitation and circumlocution.* Vocabulary usage is generally accurate. *Some errors in vocabulary choice and usage are allowed as long as they do not hinder communication.*					
Grammar (Range and Accuracy)	Shows a variety of grammar structures required for the task; shows some complex sentence structures: *- List target grammar relevant to task here.* Demonstrates a good command of grammar. *Non-systematic errors and mistakes in sentence structure may still occur. They are not impeding communication.*					
Orthography	Spelling and punctuation are reasonably accurate. *Occasional errors may still appear, especially for punctuation, and resulting from interference.*					
Space for individual comments:						

References

Baker, B. (2017). The development of EFL examination in Haiti: Collaboration and language assessment literacy development. *Language Testing, 25*, 1–25. https://doi.org/10.1177/02655322|77|6732.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*, 31–57.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education Principles Policy and Practice, 18*(3), 279–293. https://doi.org/10.1080/0969594x.2010.526585.

Brindley, G. (2001). Investigating rater consistency in competency-based language assessment. In G. B, & C. Burrows (Vol. Eds.), *Studies in immigrant English language assessment: Vol. 2*, (pp. 59–80). Sydney, Australia: Macquarie University.

Broad, R. L. (1994). "Portfolio scoring": A contradiction in terms. In L. Black (Ed.). *New directions in portfolio assessment: Reflective practice, critical theory, and large-scale scoring* (pp. 263–276). Portsmouth: Boynton/Cook (Heinemann).

Cohen, A. (1994). *Assessing language abilities in the classroom.* Boston: Heinle & Heinle.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe (2009). *Relating language examinations to the common European framework of reference for languages (CEF). A manual.* Strasbourg: Language Policy Division.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*(2), 88–115.

Figueras, N., & Noijons, J. (2009). *Linking to the CEFR levels: Research perspectives.* Arnhem: CITO, CoE, EALTA.

Gallagher, C. W., & Turley, E. D. (2012). *Our better judgment: Teacher leadership for writing assessment.* Urbana, IL: NCTE.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759–762.

Harsch, C. (2017). Noten und Kompetenzorientierung – wie geht das zusammen? [School grades and proficiency orientation – how to reconcile the two approaches?]. *Die Neueren Sprachen Jahrbuch, 5/6*, 11–22.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*, 228–250.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education, 20*(3), 281–307.

Harsch, C., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centred approach. *Language Assessment Quarterly, 8*(1), 1–34.

Harsch, C., & Seyferth, S. (2019a). Revamping the finals - Transparenz und vergleichbarkeit schriftlicher Abschlussprüfungen [Transparency and comparability of writing exams] (In print) In A. Brandt, A. Buschmann-Göbels, & C. Harsch (Eds.). *Rethinking the language learner: Paradigms - Methods - Disciplines.* Bochum: AKS Verlag: Fremdsprachen in Lehre und Forschung.

Harsch, C., & Seyferth, S. (2019b). *Evaluating a collaborative and dynamic language assessment literacy programme* (in revision).

Holzknecht, F., Kremmel, B., Konzett, C., Eberharter, K., Konrad, E., & Spöttl, C. (2018). Potentials and challenges of teacher involvement in rating scale design for high-stakes exams. In D. Xerri, & P. Vella Briffa (Eds.). *Teacher involvement in high stakes language testing* (pp. 47–66). Springer.

Kecker, G., & Eckes, T. (2010). Putting the Manual to the test: The TestDaF–CEFR linking project. In W. Martyniuk (Ed.). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft Manual* (pp. 50–79). Cambridge: Cambridge University Press.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behaviour – A longitudinal study. *Language Testing, 28*, 179–200.

Little, D., & Erickson, G. (2015). Learner identity, leaner agency and the assessment of language proficiency. *Annual Review of Applied Linguistics, 35*, 120–139.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* Frankfurt: Lang.

McNamara, T., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221–242.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.

Pearson (2018). *Global scale of English learning objectives for academic English.* Available online: https://online.flippingbook.com/view/990489. (Accessed 09.01.2019).

Seyferth, S., Lavagno, A., Kucera, P., & Harsch, C. (2018). Der GER als Grundlage zur Entwicklung sprachbergreifender Testspezifikationen [The CEFR as basis for developing test specifications acorss several languages]. In A. Brandt, A. Buschmann-Göbels, & C. Harsch (Eds.). *Der Gemeinsame Europäische Referenzrahmen für Sprachen und seine Adaption im Hochschulkontext [The Common European Framework for Languages and its adaptation fort he higher education context]* (pp. 118–144). Bochum: AKS Verlag: Fremdsprachen in Lehre und Forschung Bd. 51.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Vol. Ed.), *Studies in immigrant English language assessment: Vol. 1*, (pp. 159–189). Sydney: Macquarie University.

Studer, T., Lenz, P., & Mettler, M. (2004). Entwicklung von Instrumenten für die Evaluation von Fremdsprachenkompetenzen (Französisch/Englisch): Ziele, Kontext, Gegenstände und methodologische Aspekte des IEF-Projekts. *Revue suisse des sciences de l'éducation, 26*(3), S. 419–434.

Spöttl, C., Eberharter, K., Holzknecht, F., Kremmel, B., & Zehentner, M. (2018). Delivering reform in a high stakes context: From content-based assessment to communicative and competence-based assessment. In G. Sigott (Ed.). *Language testing in Austria: Taking stock* (pp. 219–240). Berlin: Peter Lang.

Tankó, G. (2005). The writing handbook. In J. C. Alderson (Ed.). *Into Europe.* Budapest: British Council.

Weigle, S. C. (2002). *Assessing Writing.* Cambridge: Cambridge University Press.

Wilson, M. (2008). *Constructing measures. An item-response modeling approach.* Taylor & Francis.

Claudia Harsch is a professor at the University of Bremen, specialising in language learning, teaching and assessment. Her research interests focus on language assessment, educational evaluation and measurement, intercultural communication, and the implementation of the CEFR. Claudia was the president of the European Association of Language Testing and Assessment from 2016-19.

Sibylle Seyferth was a research assistant at the University of Bremen in the field of language learning, teaching and assessment from 2016-19. Her research interests focus on language assessment of productive skills, classroom-based language assessment and teachers' language assessment literacy.

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

Claudia Harsch, Valeriia Koval, in review, in *CEFR Journal - Research and Practice*.

DO NOT PASS ON!

Abstract

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). This endeavour, however, 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 et. al. 2015; Kuiken & Vedder 2021), with a particular focus on the usability of relevant scales from the *CEFR Companion Volume* (CEFR/CV; Council of Europe 2018). 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 rating scale development reported here was part of the research project *Modelling of academic integrated linguistic competencies*, conducted at a university and a research institute in Germany (details supplied after review). The project aim was to evaluate 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, with a focus on valid rating criteria, bands, and the adaptation of existing descriptors.

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). This endeavour, however, 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 in how far relevant scales from the CEFR Companion Volume (CEFR/CV; Council of Europe 2018) 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 a combination of theory-based, descriptor-based, empirical and intuitive approaches to scale development (e.g., Chan et. al. 2015; Kuiken & 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, with the aim of offering 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 a university and a research institution in Germany (details to be supplied after review), funded by the German Research Foundation. The project is situated at 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 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 is to be used as 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, the other being member of academic faculty in English teacher education. They designed four integrated reading-into-writing tasks, two of which required students to write a

summary, 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) which was taken from introductory textbooks for freshmen in social and natural sciences. We provided detailed instructions with regard to 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. The paper here focusses on the development of the rating scale, its horizontal categories and their vertical level description. The validation and the accompanying rater training of the rating scale draft that we report here will be published elsewhere (to be supplied after review).

3. Diagnostic rating scales for integrated writing tasks

Rating scales have to be fit for their purpose (e.g., Alderson 1991; Knoch 2011); our purpose here lies on 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 task), which will be the basis for our assessment criteria (cf. section 3.3 below).

A diagnostic rating scale needs enough vertical bands or levels to inform students of strengths and weaknesses and at the same time 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., Knoch 2011). 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.

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., Knoch 2011). The wording of the descriptors should be informative for assessors (a future adaptation for learners is planned). According to North and Schneider (1989), 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 to evaluate 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 in order 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., Council of Europe 2001; Kuiken & 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 (ibid.: 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 the 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, taking the structure of the input into consideration, 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) call 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 have an impact on the integrated construct and need to be taken into consideration, as Knoch & Sitajalabhorn (2013: 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:

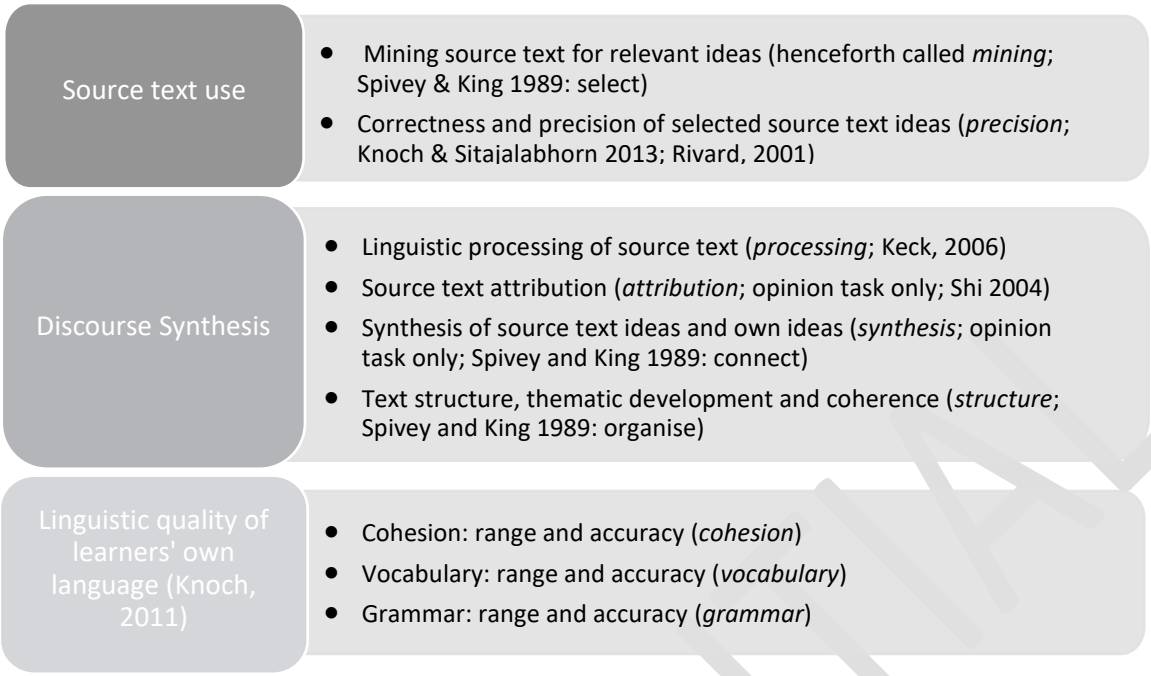

Figure 1: Diagnostic Assessment Criteria

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 test structure and thematic development under *discourse synthesis*, as it comprises 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 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*).

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 of the criteria. 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

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

Volume (CEFR-CV; Council of Europe 2018) 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., Author 1 et al., 2012; 2020; Rupp et al., 2008). Our aim was to select and where necessary adapt existing descriptors in order 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 for our criteria; we found the nine scales listed in table 1 below most informative, despite some challenges (see also Author 1 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) for differentiating the levels. 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***	***Abbre-viation***	***Challenges***
Mining	PROCESSING TEXT IN WRITING, p.112	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.62	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.63	RFIA	See RFO above, e.g., C1: academic texts vs. B1: newspaper adverts (irrelevant for our context).
Attribution	WRITTEN REPORTS AND ESSAYS, p.77	WRE	See RFO above
Cohesion	COHERENCE AND COHESION, p.142	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.132	VR	No differentiation between B1 and B1+.
Vocabulary	VOCABULARY CONTROL, p.134	VC	No +levels.
Grammar	GRAMMATICAL ACCURACY, p.133	GA	Descriptors for range of structure not consistent and only mentioned at B1 and B2.
Vocabulary Grammar	ORTHOGRAPHIC CONTROL, p.137	OC	No +levels.

Like in similar scale development projects (e.g., Author 1 et al. 2012, Author 1 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 to 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 also would 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.128	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.129	Levels differentiated by varying domains, target audiences or topics, which is not relevant for our context.
THEMATIC DEVELOPMENT, p.141	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).
Appendix 4 – Manual Table C4: Written Assessment Grid, p.173	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.

Appendix 9 – supplementary descriptors, scale ADAPTING LANGUAGE, p.232	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***	***Abbrev iation***	***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): APPENDIX D – RATING SCALES FOR WRITING TASKS, levels B1-C1, criteria organization and grammar, (pp.149-155).	IQB	CEFR-based rating scale, validated (Author 1 et al., 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 is grouped in the first criterion. While the different aspects are coherently defined on six levels, the levels are not aligned to 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

After piloting the integrated tasks, we analysed the collected scripts (84 texts produced by students, 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 were 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, other 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	- scripts	- Shi 2004 - Keck 2006 - scripts	- *CEFR* - *scripts*	- *scripts*	- CEFR - IQB - Li, 2014	- CEFR - IQB	- CEFR - GSE	-CEFR - 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 not 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 rate 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 the development of 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 the stance they took). With regard to differentiating levels 4 and 5, the experts suggested to add for Level 4 „may contain some irrelevant ideas“. They also suggested to add the depth of understanding 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, which is described and examined elsewhere (to be supplied after review). Another recommendation was to add a special code for cases where writers only used their own ideas (and hence no paraphrasing could take place). The criterion *Attribution* was perceived as clearly worded and feasible, while for criterion *Synthesis*, the experts recommended to specify 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. 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. Finding suitable descriptors to describe these criteria proved to be more challenging. While the CEFR-CV provides a rich source of scales from which to choose relevant descriptions, not all scales were feasible for our context and construct elements. This holds particularly true for those scales that use domains, target audiences or topics which were not relevant for our context. 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 Author 1 et al. (2012, 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 Author 1 et al. 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.

Acknowledgement

Our thank 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 colleague (name to be provided after review) for contributing to the qualitative script analyses. Our heartfelt thank goes to our colleague (name to be provided) 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.

References

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

Author 1 et al. 2012

Author 1 et al. 2020

Chan, S.H., Inoue, C., & Taylor, L. (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. 2018. *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg: Language Policy Division. https:// rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989 (accessed 27 November 2019).

Cumming, A. (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), https://www.cuny.edu/wp-content/uploads/sites/4/page-assets/academics/testing/CATWInformationforStudentsandpracticeweb.pdf (1.7.2022).

ETS (n.d.). TOEFL Writing Rubrics, https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf (1.7.2022).

IELTS. 2013. *IELTS TASK 1 Writing band descriptors* (public version). https://www.ielts.org/-/media/pdfs/writing-band-descriptors-task-1.ashx?la=en (accessed 1.7.2022).

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

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

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

Kuiken, F., & Vedder, I. (2021). Scoring Approaches: Scales/Rubrics. In P. Winke, & T. Brunfaut (Eds.), The Routledge Handbook of Second Language Testing (1st ed.). Routledge. DOI https://doi.org/10.4324/9781351034784

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

Li, J. & Wang, Q. (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). DOI https://doi.org/10.1186/s40862-021-00113-6

McNamara, T., Morton, J., Storch, N., & Thompson, C. (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, DOI: 10.1080/15434303.2017.140542

North, B., & Schneider, G. (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*. http://www.b-li.ir/ar/06CEFR/4.GSE_LO_Academic_English.pdf (accessed 12 February 2020).

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

Plakans, L. & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing 31*, 98-112.

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

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

Sawaki, Y., Quinlan, T., & Lee, Y. (2013) Understanding Learner Strengths and Weaknesses: Assessing Performance on an Integrated Writing Task, *Language Assessment Quarterly*, 10:1, 73-95, DOI: 10.1080/15434303.2011.633305

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

Spivey, N. N., & King, J. R. (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. https://www.trinitycollege.com/resource/?id=7468 (1.7.2022).

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 sections for the task at hand (RFO: B2+).	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 relationships between ideas (CC: B2+).	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 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).*

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 - *some irrelevant details*	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 mark clearly the relationships between ideas.	-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 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 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 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 (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 main ideas selected. -Includes irrelevant details or irrelevant own ideas (summary).	*Majority of selected ideas may be misinterpreted.*	Near copy: -Major instances of lifting from ST (usually without referencing). Add Code 0 (not applicable): writes only own ideas	-Generally difficult for reader to distinguish between ST and own ideas. -Ideas from source text are generally not attributed to ST.	-No clear stance. -Barely relies on ST for argumentation, offering own (mis-) interpretation of the topic. -Or merely summarizes ST, not adding relevant own stance.	-Orders a series of shorter discrete elements into a linear sequence of points. Structure/thematic development follows ST, although not appropriate for the task. Or: *may lack a logical order appropriate for the task.* -Paragraphs usually not appropriate (if used at all).	-Links discrete elements using a limited number of cohesive devices. -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.*	-Sufficient range of vocabulary, with some circumlocutions, repetitions. *-May show some instances of inappropriate vocabulary use.* *Major errors may occur when expressing more complex thoughts.* -Spelling is accurate enough to be followed most of the time.	-Uses a repertoire of frequently used "routines" and patterns associated with more predictable situations reasonably accurately. *-May attempt complex patterns but generally unsuccessfully.* -Punctuation is accurate enough to be followed most of the time.