https://doi.org/10.17323/jle.2025.27560

Assessing Academic and Disciplinary Literacies: Rubric Validation to Mjeasure Argumentation, Comparison and Source-Based Writing Skills

Roberto Arias-Hermoso, Eneritz Garro Larrañaga, Ainara Imaz Agirre

Mondragon Unibertsitatea, Spain

ABSTRACT

Introduction: Over the last two decades, research has suggested that academic and disciplinary literacies (ADLs) are key to integrating content and language, as language is used to express content knowledge, often through Cognitive Discourse Functions, discourse patterns that respond to cognitive actions in formal education contexts. Nevertheless, systematic assessment tools are required to assess student production of ADLs.

Purpose: This validation study focuses on developing an analytic rubric to measure three ADL skills, consisting of three dimensions. Two of these dimensions are CDFs: students' skills to *argue* and *compare.* A third rubric measures an additional academic skill, students' ability to write from sources. The rubrics are designed to capture cross-disciplinary and multilingual productions of these skills, therefore being applicable for various disciplines (history, science, mathematics, among others…) and languages.

Method: A five-step validation process based on expert judgement was used, involving 13 international experts. They quantitatively and qualitatively evaluated the proposed rubrics based on pertinence, conceptual clarity, coherence, and relevance across two iterative rounds. Quantitative descriptive statistics and agreement indices were used, in addition to thematic analysis of qualitative feedback.

Results: The quality of the rubric showed clear progression between validation rounds. In the first version, several issues were raised by the experts–most criteria showed weak agreement and low means. After revisions, there was a substantial improvement in the second version, with 83 % of the criteria reaching strong or acceptable agreement. Qualitative feedback highlighted the need for precise and multidimensional operationalisations in each ADL dimension. To illustrate the application of the rubrics, multilingual student samples are provided.

Conclusion: The rubrics offer the first steps towards systematising ADL assessment by combining qualitative and quantitative feedback from 13 experts, underscoring the importance of expert input in advancing assessment practices. The study has theoretical and practical contributions: it highlights the multidimensional nature of ADLs and provides an adaptable rubric that can be used across disciplines, languages and educational contexts. This study focused on the validation process; therefore, empirical use of the rubric is still required. Future research should apply the rubrics to large-scale corpora and complement the expert-based validation with psychometric approaches.

KEYWORDS

cognitive discourse functions; writing assessment; disciplinary literacies; source-based writing

Citation: Arias-Hermoso, R., Garro Larrañaga, E., & Imaz Agirre, A. (2025). Assessing academic and disciplinary literacies: Rubric validation to measure argumentation, comparison and source-based writing skills. *Journal of Language and Education, 11*(3), 60-75. https://doi.org/10.17323/jle.2025.27560

Correspondence:
Roberto Arias-Hermoso,
rarias@mondragon.edu

Received: July 03, 2025
Accepted: September 18, 2025
Published: September 30, 2025

INTRODUCTION

Mastering academic language has been defined as a critical factor for educational success (Lorenzo, 2023), as it is through language that learners construct, communicate and are assessed on school-related knowledge (Schleppegrell, 2004). In fact, every aspect of the teaching and learning process entails some type of language: building new knowledge through classroom interaction (Nikula et al., 2024), answering exam prompts (Breeze & Dafouz, 2017), or teaching

materials (Lersundi, 2022). All of these require students to *do* something with language, shaping their thoughts and knowledge. It is, therefore, essential to understand that language processes cannot be isolated from content learning. Consequently, implementing and considering educational interventions that foster academic language skills in integration with content is necessary (Imaz Agirre et al., 2025).

The disciplinary dimension of academic literacies has been highlighted, since academic literacy is heavily influenced by "the shared cultural values and communicative repertoires within the disciplines" (Li, 2022: 2). In this regard, recent research on content and language integration learning (CLIL) and bilingual education has shifted its attention towards the disciplinary aspect of academic language, i.e., disciplinary or subject-specific literacies (Dalton-Puffer et al., 2024). Despite nuances in its definitions and operationalisations, disciplinary literacy has been defined as the set of conceptual and linguistic skills needed to construct and communicate subject-specific knowledge (Moje, 2015; Shanahan & Shanahan, 2008; Dalton-Puffer et al., 2024). Disciplinary literacies thus extend academic language skills (Shanahan & Shanahan, 2008) and provide a pedagogical recontextualization of professional expert domains for school purposes (Fang & Coatoam, 2013). In this article, we focus on Academic and Disciplinary Literacies (ADLs), which include both aspects related to more than one discipline, and, therefore, cross-disciplinary, as well as specific disciplinary practices.

As a way of connecting the discourse practices that emerge in multiple disciplines, but at the same time considering specific disciplinary aspects, Dalton-Puffer (2013) proposed the construct of Cognitive Discourse Functions (CDFs hereafter), which has been suggested as central to understanding ADLs, since it perfectly captures the interconnectedness between content, language and literacies (Morton, 2020). CDFs represent verbalisations of cognitive patterns common to educational contexts, such as *defining, exploring* or *explaining*. These patterns require specific discourse and lexicogrammatical schemata for their realisation (Dalton-Puffer & Bauer-Marschallinger 2019), and are grouped into seven major categories: categorise, define, describe, explain, explore, evaluate and report. The CDF construct has been gaining international attention since its conception, and several studies have used it to examine classroom interaction (Lersundi, 2022), written exam responses (Breeze & Dafouz, 2017) or pedagogical interventions (Rieder-Marschallinger, 2024). A central aspect of CDFs is their elicitation in CDF-tasks (Gerns, 2023a), in fact, learners are often asked to perform CDF-tasks, defined as processing subject contents and their verbalisation through corresponding language forms (Gerns, 2023a). Most disciplinary productive tasks elicit some kind of CDF, such as defining what the ozone layer is, evaluating how cheques and bills affected trade or classifying triangles according to their angles (all examples from Lorenzo et al., 2024).

CDFs, therefore, represent both the cognitive (content) and discourse (language) resources needed by students to perform certain ADL tasks. In these tasks, students often respond based on their previous knowledge of the content matter, normally acquired through school settings. Nevertheless, many disciplines have emphasised the importance of a crucial academic skill: using discipline-related sources to construct knowledge. For example, Sendur et al. (2021) or Steiss et al. (2024) mention that using and corroborating sources is crucial for historical literacy. Research on writing from sources, or source-based writing (SBW), has provided ample evidence of the types of tasks that can be performed after using sources, with argumentative and summary tasks being the most frequently used ones (Chan & May, 2022). These fall within the scope of CDFs, as *arguing* and *summarising* are indeed part of the CDF construct. Although the connection between CDFs and SBW has yet to be well established in the literature, both represent important academic literacy skills that are activated in educational contexts.

The usefulness of the CDF construct for research and teaching has been emphasised, and Lorenzo et al. (2024) claim that the construct offers great opportunities for ADL assessment. Nevertheless, while various approaches have been proposed to assess ADLs, existing tools may be limited and fail to capture the multidimensional nature of each CDF. In particular, there is a lack of validated analytical–not holistic–rubrics that are adaptable across languages and disciplines. This article, therefore, addresses this gap by reporting the validation process of an analytic rubric focusing on two specific CDFs (*argue* and *compare*) and the ability to write from sources. The following sections will summarise research on ADL assessment, as well as provide contextualisation on how these three skills have previously been operationalised for assessment purposes, therefore providing evidence of the most important components of the three dimensions of the rubric presented here.

LITERATURE REVIEW

Assessing Academic and Disciplinary Literacies

Despite the fact that ADL assessment still needs to be further explored and developed to create systematic assessment tools, previous research has used diverse approaches. These can be grouped into three broad categories: studies using component-based assessment, those using comparative judgement and those using rubrics or scales.

Component-based assessment of ADLs has focused on identifying certain features of ADLs and CDFs in students' production by, normally, employing conceptual maps or Systemic Functional Linguistics. For instance, the work by Connolly (2019) analysed secondary education students' scientific explanations. Her component-based assessment

model included conceptual sophistication, addressing reasoning and validity, and communicative sophistication, addressing global and local cohesion and subject specificity. Similarly, other studies have proposed conceptual frameworks to analyse specific CDFs, such as *defining* (Nashaat-Sobhy & Llinares, 2023), *arguing* (Garro Larrañaga et al., 2024), *reporting* (Roca de Larios et al., 2023), *comparing* (Gerns, 2023; Evnitskaya & Dalton-Puffer, 2023, 2024) or *evaluating* (Whittaker & McCabe, 2023), most of them based on Systemic Functional Linguistics (see Halliday & Matthiessen 2014). The studies, overall, suggest that conceptual frameworks and maps can be beneficial to understanding how learners perform in CDF-tasks (Gerns, 2023). Nevertheless, while valuable for research purposes, these frameworks can often be too complex for classroom use.

A second group of studies has used comparative judgement techniques (Lesterhuis et al., 2017) to explore the main assessment criteria highlighted by teachers when assessing learners' ADL production. Comparative judgement uses algorithms that, after a series of pair comparisons of texts by a number of judges, calculate scores based on the number of wins and losses. Llinares et al. (2024) found that when reflecting on comparative judgement procedures to assess ADLs, content and language teachers' language awareness was fostered, and it helped them articulate their thoughts on CLIL. Similarly, Nashaat-Sobhy and Morton (2024) and Morton (2022) identified the most important characteristics when teachers evaluated learner writing, and proposed that content quantity and quality, language function and form, and the integration of all of these played an important role in assessing ADL writing. Similarly, the study by Meneses et al. (2023) found that the comparative judgement scores given by the teachers were highly correlated with rubric and analytic scores. Although it provides promising opportunities, comparative judgement with teachers has yet to yield systematic descriptors for specific CDFs, following Llinares and Morton's (2024) work with the CDF *explore*, moving from teacher awareness to specific tools.

A final group of studies focusing on ADL has used rubrics, sometimes in combination with the other two strategies mentioned above, such as the studies by Gerns (2023a) and Connolly (2019) using component-based assessment and a rubric, or Meneses et al. (2023), using both comparative judgement and an analytical rubric. The usefulness of rubrics and scales has been highlighted in the literature, as they have frequently been used in educational research and/or language testing (Kuiken & Vedder, 2020). Rubrics, in fact, can provide a detailed account of a learner's performance levels in certain tasks, as they normally provide descriptors or indicators reflecting students' performance. In the field of CDF research, the use of rubrics can prove useful for both teachers and researchers to capture performance in students' writing (Granados & Lorenzo, 2024; Lorenzo et al., 2024). Different studies examining CDF productions have proposed specific rubrics for specific CDF-tasks, such as de-Boer (2020) for online interaction contexts, del Pozo and Llinares (2021) and del Pozo (2024) for history writing, or Bauer-Marschallinger (2022; Rieder-Marschallinger, 2024) for historical literacy. Although not necessarily from a CDF perspective, other rubrics have also been used to assess the skills under study in this article, i.e., rubrics to measure argumentation (e.g., Qin & Karabacak, 2010) or SBW skills (Uludag & McDonough, 2022).

Many of these rubrics, whether CDF-related or not, however, normally consider the skill as a whole. Lorenzo et al. (2024), for example, provide a list of CEFR-based descriptors for B1 and B2 levels for the 7 CDFs in history, mathematics and science. While the disciplinary differences capture important nuances of each subject, the descriptors measure each CDF as a whole unit holistically (e.g., *defining* at B1 in mathematics) and may fall short when observing specific aspects that can lead to nuanced interpretations. This is a critical gap in CDF assessment for research, as ADLs are inherently multilayered and multidimensional (Nikula et al., 2024; Dalton-Puffer et al., 2024). Therefore, reducing CDF assessment to a single descriptor may fall short on capturing nuances of student performance. Consequently, research has called (del Pozo, 2024) for the use of analytical scales that consider various aspects of the same dimension, which improve construct validity (Paltridge & Phakiti, 2015). In the following paragraphs, we define and operationalise how the three main dimensions under study (the CDF *argue*, CDF *compare* and source-based writing) have been assessed previously in the literature.

The CDF Argue

Arguing is a sub-form of the CDF category *evaluate*, which Dalton-Puffer (2013: 234) defined as "I tell you what my position is vis a vis X". Arguing goes beyond expressing one's position, as it focuses on defending that position using evidence (Dalton-Puffer, 2016; Toulmin, 1958). Argumentation could thus be understood as the ability to use data or evidence to support one's opinion (Jiménez-Aleixandre & Erduran, 2007), and has been highlighted as a central element of education overall (Pylonitis & Meyer, 2024), and both in science (Polias, 2016; Jiménez-Aleixandre & Erduran, 2007) and history education (Lorenzo, 2017; Coffin, 2009; Sendur et al., 2021).

According to Crossley et al. (2022), learner argumentation has traditionally been analysed using Toulmin's model (1958). The model includes basic and complementary elements: the claim (the writer's opinion), data supporting the claim, warrants justifying how the data supports the claim, or rebuttals, refuting exceptions in which the claim might not be true. Toulmin's model, however, has received certain criticism (see for example Liu & Stapleton, 2014), as it does not include counterarguments, which are an essential part of the argumentative structure. Because of that, adaptations and modifications based on Toulmin's model have

been proposed (Sampson & Clark, 2008; Qin & Karabacak, 2010; Liu & Stapleton, 2014; Crossley et al. 2022). Using Toulmin's model as a starting point, Qin and Karabacak (2010) provide a holistic rubric from 1-5 measuring a writer's ability as a whole. Other proposals (e.g., Stapleton and Wu, 2015; Allagui, 2019; Jackson, 2024) specified different levels for each of the argumentation elements in Toulmin's model. Similarly, Sendur et al. (2021) proposed a scale for CLIL History, including different levels for the claim and evidence. All in all, these studies point out that successful argumentation quality is characterised by a clear claim supported by empirical data, with explicit justifications and rebuttals. These rubrics, however, often focus on argumentation skills at the textual level in higher education, when students typically have a greater understanding of argumentative writing. In addition, they frequently rely on vague descriptors (e.g., 'the claim is clear' vs. 'very clear'), which lacks specificity and operational precision.

CDF Compare

Comparing falls within the scope of the CDF category *categorise*, defined as "I tell you how we can cut up the world according to certain ideas" (Dalton-Puffer, 2013: 234). Comparing is a crucial step to categorise – it is fundamental to establish how two elements are equal and/or different according to certain criteria to be organised into categories (Evnitskaya & Dalton-Puffer, 2023). Comparing, therefore, is much about identifying similarities and differences between two or more elements, e.g., how are carnivores and herbivores different from each other? (Gerns, 2023a, 2023b). Despite the importance of comparing to build new knowledge by creating (dis)connections between concepts (Chen & Zhou, 2022), not much research has focused on learners' comparison skills. Nevertheless, it has been proposed as a central function in science (Gerns, 2023), and important for history too (Lorenzo et al., 2024).

The identification of a good comparison, however, is not as straightforward as argumentation, since comparison-and-contrast has normally been addressed as a genre move in scientific articles (see for example Chen & Zhou, 2022). In the domain of ADLs, however, research on comparing has recently emerged (Gerns, 2023a, 2023b; Evnitskaya & Dalton-Puffer, 2023, 2024). Based on previous proposals (for example, Trimble 1985 for science), the studies suggest that three main elements constitute the structure of comparison: what is being compared (the comparative item), the criteria used to establish similarity/difference, and the textual structure of the comparison. To our knowledge, only Gerns (2023a) has proposed a rubric to assess students' comparison skills. The rubric consists of three quality criteria: completeness, precision and explicitness. According to her rubric, students' comparisons receive higher quality scores when, in general, they write more balanced and complete comparisons (including both similarities and differences in a balanced way), the criteria are explicit and well-justified, and the structure is parallel and clear. However, research on *compare* as a CDF has recently begun to provide operational instruments (Gerns, 2023a) and do often not focus on the quality of specific elements, i.e., comparative items or criteria.

Source-Based Writing

The final aspect of interest is source-based writing (SBW), which is defined as the ability to use input sources to produce written compositions (Cheong et al., 2021). Although the connection between SBW and CDFs has not been explored in depth, they can easily be linked. In fact, students are often asked to produce a CDF-based composition in a SBW task. According to Chan and May (2022), arguing and summarising (falling under *report*) would be the most frequent functions activated in SBW tasks. In SBW tasks, learners use diverse inputs to construct compositions. These inputs, or sources, could be multimodal resources, which has been highlighted as one of the pillars of ADLs (Nikula et al., 2024), as students are exposed to various semiotic modes in the learning process, such as texts, audiovisual materials or graphs, some of them subject-specific (a timeline in History, a formula in Science or Mathematics).

Different aspects of SBW have been studied in the literature, particularly the ability of students to 1) use information from the sources, and 2) their ability to transform the source using their own words (e.g., Arias-Hermoso et al., 2024; Kato, 2018; Plakans & Gebril, 2013; van Weijen et al., 2019). Rubrics that consider different aspects of SBW have been previously proposed in the literature. The previously mentioned scale by Sendur et al. (2021) for CLIL History, in addition to the two items focused on argumentation, also measured students' ability to evaluate the sources, corroborate information and use the sources to contextualise the historical period. The rubric proposed by Uludag and McDonough (2022), on the other hand, focused on content (addressing the argumentative prompt correctly), organisation and language use, with an additional specific item to assess *source use*. This item considered both students' comprehension and paraphrasing of the source, as well as citing (Harsh et al., 2024). While some rubrics for SBW exist, they tend to focus on content and organisation–general aspects of writing–and citation practices. These overlook other crucial aspects of SBW, such as transforming sources or integrating students' own knowledge alongside information from the sources. The latter has been shown a critical aspect of SBW quality, closely related to creativity (Arias-Hermoso et al., 2024; Cheong et al., 2021), but has not been systematically included in previous rubrics.

Despite the valuable contributions of previous studies assessing ADLs, as has been mentioned, existing tools that focus on CDF assessment remain limited. Consequently, previous research (e.g., del Pozo, 2024) has called for more systematic and multidimensional approaches that can cap-

ture nuanced patterns in ADL student productions. Furthermore, many of the existing rubrics or component-based frameworks are tailored to productions written in English, ignoring the multilingual nature of ADLs. Together, these limitations underscore the need for validated and multidimensional rubrics that can assess specific CDFs and other ADL-related skills across languages and disciplines. Therefore, this study has a main research objective (RO) and two sub-objectives:

RO1: To validate an analytic rubric to assess students written production of the CDF *argue*, the CDF *compare* and their source-based writing skills.
RO1.1: To refine the rubric based on ADL expert judgement, ensuring clarity and coherence.
RO1.2: To illustrate the rubrics' applicability across languages and disciplines.

METHOD

Context and Intended Corpus

The rubric was designed to be applied in a multilingual disciplinary source-based writing corpus to better understand students' ADL development. As has been said, the present paper focuses exclusively on the validation of the rubric through expert judgement. While it does not report empirical analyses of the corpus, it serves two purposes. First, it aims to guide the design of rubric components to ensure applicability across languages and disciplines, and second, it aims to provide illustrative examples from the corpus of how the final rubric can be used. The empirical findings related to students' development of ADLs using the rubrics presented here are reported in Arias-Hermoso et al. (2025a) and Arias-Hermoso et al. (2025b), with a focus on either the language of instruction or the three languages.

As background, the intended corpus consists of texts written by 535 students across secondary education (ages 12-16) in a Basque-immersion model[1]. In this model, students typically receive 26 hours a week of exposure to Basque (in language and content classes) and 3 hours to both Spanish and English. In these models, the majority of students have Basque and/or Spanish as their first language, with the other being a second language, and English being a third language. Due to extensive schooling and to their status as the languages of the context, students in secondary education tend to have a good command of both Basque and Spanish (around B2), while English proficiency is usually lower (around A2).

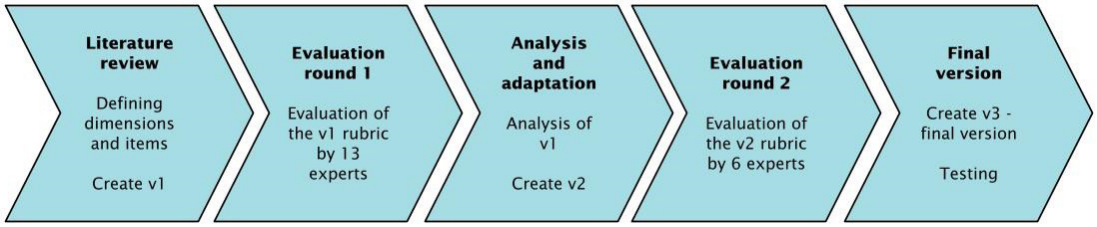
The intended corpus includes students' texts in their three languages of the curriculum (Basque, Spanish and English) and two disciplines (Science and History). The corpus is based on a SBW task in which students have to argue in favour of one of two subject-specific elements by comparing them, using information from the given sources. In the field of Science, the topic is nutrition, and learners have to compare the nutritional values and characteristics of two foods (chicken and chocolate), to establish which is healthier. In History, students are expected to compare the Middle Ages and today's society, using information about politics, economy, education and health. The compositions are written as argumentative blog entries for the school blog and were expected to be around 150 words. All participants and their legal guardians were informed about the nature of the research, and signed informed consent was collected before participation.

It is important to mention that, because of the nature of the task (cross-disciplinary and multilingual), the rubrics were created to be applied to different languages and disciplines. Furthermore, the rubric was planned to measure the three abilities in this particular type of task. Nevertheless, the three dimensions could also be used independently from each other, by, for instance, assessing source-based writing skills alone in a non-argumentative or non-comparative task.

The Validation Process

Various techniques have been used in educational research to address construct and content validity. For the purposes of this study, an expert judgment procedure was implemented. This procedure has been previously used in educational research (e.g., Lázaro Cantabrana & Gisbert Cervera, 2015) and entails the collection of experts' evaluations and feed-

Figure 1
Validation Process



[1] For a detailed description of the participants, the data collection procedures and empirical results, the reader is encouraged to read Arias-Hermoso et al. (2025a, 2025b), which present the systematic and empirical aspect of the developmental study.

back on the instruments, and then modifying them accordingly until quality standards are met (Landeta & Lertxundi, 2023). As has been mentioned, the main objective of the rubric of this study is to capture quality-related differences in three dimensions (argumentation skills, comparison skills and SBW) by considering multiple components of each. As can be seen in **Figure 1,** the validation process was carried out in five differentiated steps, which will be described in detail in this section.

Step 1: Literature Review and Creating V1

In the first step of the validation process, a literature review on each of the dimensions was conducted. Special focus was placed on identifying the key components or elements essential for successful performance within each dimension. In other words, we reviewed the literature focusing on answering the following questions: 'What are the essential features of *argumentation, comparison* and *writing from sources*? Which of these features signal better or more advanced performance?'. To answer these questions, we also examined already existing rubrics that measured the same dimensions (see Literature Review above). Due to the multidimensional nature of ADLs (Dalton-Puffer et al., 2024; Nikula et al., 2024), it is essential to consider how students perform at different components of the same dimension, as strongly highlighted by del Pozo (2024).

Following the suggestions by Vercellotti and McCormick (2021), the first step then consisted of identifying separate categories within each dimension. Therefore, after the literature review, the main components of each dimension were identified, and the first version of the rubric (V1) was created. The dimensions were called as follows: argumentation skills (D1), comparison skills (D2) and source-based writing (D3). Each dimension was then divided into items, representing the different elements or components expected for successful performance in said dimensions. Both D1 and D2 were composed of 4 items and D3 of 2. Then, the performance levels for each item (from 0, the lowest, to 3, the highest) were described based on the literature review.

Step 2: Expert Evaluation: Round 1

The second step consisted of creating the evaluation matrix and recruiting the experts for the evaluation process. An Excel sheet with the instructions and the evaluation matrix for each dimension was designed. Experts were instructed to rate each item from the rubric against 4 assessment criteria using a Likert scale with indicators from 1 to 5, as well as adding possible qualitative suggestions and comments for each of the items, or the dimension in general. The assessment criteria included the following four aspects:

(1) Pertinence (PE): the extent to which the item's formulation corresponds to what it measures.
(2) Conceptual clarity (CC): the item is accurately formulated and easy to comprehend, avoiding confusion or contradictions.
(3) Coherence (CO): the item has a logical connection with the dimension or indicator that it measures.
(4) Relevance (RE): the item is essential or important, i.e., it must be included.

In addition to an explanation of how to use the evaluation matrix, the document included a detailed description of the task that was going to be assessed with the scale. Subsequently, 15 international expert raters were contacted, out of which 13 accepted doing the evaluation process. All 13 experts held a PhD, had ample experience in language education and/or applied linguistics, and were familiar with the CDF construct and ADLs. In fact, most experts had previously published articles on CDFs and/or ADLs in peer-reviewed journals, and some also had extensive experience with either Science or History education. This decision was based on the importance of including both content and language experts' insights to improve the scale. They were given 2 months to complete the evaluation. All experts were informed about the nature of the validation process and written informed consent was received.

Step 3: Feedback Integration and Adaptation (V2)

After receiving the ratings from the expert judges, an analysis of their responses was conducted, both quantitatively and qualitatively. In a quantitative approach, descriptive statistics (means, standard deviations (SD), mode, minimum and maximum values) were calculated for each item and criterion. While Cronbach's alpha is suitable for within-dimension consistency, and ICC assume multiple targets (e.g., texts, for example), these are not suitable for this design. Other models (i.e., Rasch measurements) are often employed to validate rubrics, but they require multiple raters scoring multiple texts. In this case, raters did not *use* the rubrics but rather evaluated them (meta-assessment). Therefore, inter-rater agreement was calculated with dispersion measures: strong agreement was considered when the SD was below 0.5, acceptable between 0.51 and 0.8, and weak below 0.81. These measures were used in combination with the qualitative feedback to improve the rubric (see Results). Despite most items reaching an average score above 4 (M of all items and criteria = 4.32, SD = 0.51), especially for D1 and D3, some important issues with the items were raised in the qualitative feedback. The qualitative feedback analysed using thematic analysis, i.e., identifying common patterns in the comments. The common patterns were grouped for each item. This step was crucial to improve the rubric, and, accordingly, some important changes were made. Some items were removed, others were added, and some were divided into two different items. Most were reformulated following the feedback. All of the changes were summarised in a written report that compared V1 with the modified second version (V2).

Step 4: Expert Evaluation: Round 2

The report was sent back to the 13 raters, and they were asked to complete a second round of evaluation to assess whether the adaptations were sufficient. Out of the 13 original judges, 6 completed a second round of evaluation. The procedure was identical to the first round of evaluation: raters had to complete an evaluation matrix for each item assessing the 4 criteria. After this round, the feedback was considered to be sufficiently positive, with the vast majority of items reaching average values above 4.5, and all above 4. Some minimal changes were made after this round of suggestions.

Step 5: Feedback Integration and Adaptation (V3)

The final feedback was integrated and a final (V3) version was created. This version was piloted with 12 texts by the three authors until reaching a total agreement.

RESULTS

In this section, we present the results divided into two main sections: first, we focus on the development and validation process of the rubric by including a summary of the quantitative and qualitative feedback from the experts. Subsequently, we will highlight the main changes to each dimension. The second section will illustrate the final version of the rubric, supported by real student examples in English to facilitate reading the examples to an international audience, as well as instructions for research and practice.

Integration of Expert Judges' Evaluation

As can be seen in Table 1, there was a quantitative improvement from V1 to V2, particularly in D2, which presented the lowest scores. Quantitative results are based on the expert rating matrix, in which each item was judged on the four criteria using a 5-point Likert scale. In V1, following the SD threshold to calculate inter-rater agreement (i.e., SD ≤ 0.50 strong, 0.51-0.80 acceptable, > 0.81 weak), only 8 criteria (18.6 %) showed strong agreement, 9 (20.9%) acceptable, and 26 (60.5%) weak. In particular, items related to conceptual clarity tended to generate higher variability across judges (e.g., stance-taking [1.09]). After the revisions, not only did the averages improve, but also agreement: 36 criteria (76.6 %) reached strong agreement, 3 (6.4 %) acceptable and only 8 (17 %) weak. All items with weak agreement had a SD between 0.82-0.85, therefore being almost acceptable (see Supplementary Materials for complete descriptive and interrater agreement statistics[2]).

Nevertheless, the progression demonstrates that the iterative validation process, guided by expert feedback, significantly improved the rubric's clarity and coherence. Some general changes were made throughout all items of the rubric, such as the formulation patterns. V1 included different formulations for various items, such as 'the writer has presented' or 'the text presents', with occasional use of the

Table 1

Means for each Item in V1 and V2, Divided by the Assessment Criteria

Version 1 (13 raters)	PE	CC	CO	RE	Version 2 (6 raters)	PE	CL	CC	RE
D1. Argumentation skills	Sufficiency = 100 %				**D1. Argumentation skills**	Sufficiency = 100 %			
Stance-taking	4.69	4.23	4.62	5	Stance-taking	5	4.67	5	5
Evidencing	4.54	4.15	4.62	4.85	Reasoning	5	4.67	5	5
Opposition	4.69	4.31	4.46	5	Counterargumentation and rebuttal	5	5	5	5
Use of argumentative language	4.46	4.23	4.5	4.42	Language for arguing	5	4.5	5	5
D2. Comparison skills	Sufficiency = 90 %				**D2. Comparison skills**	Sufficiency = 100 %			
Comparison	4.45	3	4.55	4.82	*Comparative structure*	*4.67*	*4.83*	*5*	*5*
Comparees	3.85	2.85	3.85	3.54	Comparative items	4.83	4.67	4.83	5
Criterion	4.5	3.17	4.17	4.75	Comparative criteria	4.33	4.5	4.83	4.83
Use of comparative language	4.33	3.92	4.42	4.67	Language for comparing	5	4.6	5	5
D3. Source-based writing	Sufficiency = 70 %				**D3. Source-based writing**	Sufficiency = 100 %			
Use of source information	4.23	4.15	4.31	4.86	Use and comprehension of sources	5	4.60	5	5
Sourcing (degree of transformation)	4.54	3.69	4.62	5	Degree of transformation	5	4.40	5	5
					Use of original ideas	*5*	*4.60*	*5*	*5*

Note. Italics represent items that were added after the feedback, as they were not included in V1.

[2] Supplementary Materials available here: https://doi.org/10.6084/m9.figshare.30069613.v1

passive voice. In order to avoid misunderstanding and to ensure consistency, all items were reformulated to 'The writer presents/elaborates…'. An important issue present in D1 and D2 was raised by some of the judges – the fourth item within these dimensions referred to *the use of argumentative/ comparative language*. As has been said in the Introduction, it is true that language (form) cannot be separated from the production of discourse-functional patterns or elements related to the item. For example, stance-taking requires using certain lexicogrammar choices such as *I think* or *in my opinion* (Liu & Stapleton, 2014). Some judges suggested, therefore, merging this category into the corresponding item, instead of a single dimension. However, as also suggested by other judges, students might perform well at a functional level (be able to reason their claim with arguments) but fail to use accurate, appropriate or varied language forms, an issue also noted by Bauer-Marschallinger (2022), particularly in the context of additional (L2/L3) language learners. Therefore, we understand these two items as the lexicogrammatical coating of CDF production. Consequently, the first three items in D1 and D2 measure functional-structural aspects of CDFs, while the formal-linguistic aspects are measured by the fourth items.

Changes to D1: Argumentation Skills

The main changes carried out to D1 after the feedback are summarised here. V1 considered restating the claim as the highest value, but it was deemed unnecessary due to text length. In addition, inconsistent claims were considered part of 1, but they were moved to 0. The other three items' names were also changed: *evidencing* became *reasoning*, a common term used in both scientific and historical literacy research (e.g., Coffin, 2009). In this item, the previous formulation used 'evidence' to refer to both data and warrants, but following main research on argumentation (Toulmin, 1958; Crossley et al., 2022; Liu & Stapleton, 2014; Qin & Karabacak, 2010), the term 'evidence' was erased, and only data and warrants were included. *Opposition* was considered too broad and not entirely consistent with what the item intended to measure, and thus it was renamed to *counterargumentation and rebuttal*, following research on argumentation (Toulmin, 1958; Crossley et al., 2022; Liu & Stapleton, 2014; Qin & Karabacak, 2010). V1 did not establish very clear differences between levels 2 and 3, and it was modified to include a more explicit use of rebuttals. *Use of argumentative language* was changed to *language for arguing*, so as to include a wide array of lexicogrammatical resources used to produce arguments, including morphosyntactic, lexical, grammatical and textual level resources.

Changes to D2: Comparison Skills

This dimension experienced the greatest changes from V1 to the final version. The initial version included an item called *comparison*, which purposefully included the comparative items and the criteria, but a split version was recommended by all of the judges. Therefore, this item was erased, and the items were considered separately: *comparees* was changed to *comparative items* (following Gerns, 2023; Chen & Zhou, 2022), and *criterion* to *comparative criteria*. One of the experts suggested that "equally elaborating on differences and similarities among the two compared items is a crucial aspect of comparison", therefore, this was added as the highest level of *comparative items*. V1 included the basis of classification (following Trimble, 1985) as the main comparative criterion, however, this was reformulated in the definition to 'basis of comparison' in order to better capture the item's focus (Gerns, 2023; Evnitskaya & Dalton-Puffer, 2023, 2024).

As suggested by some judges, an additional dimension was added to examine the structure of the comparisons, *comparative structure*, as using a parallel structure has been highlighted as a crucial move in comparing (Gerns, 2023; Smith, 2019). The structures include the point-by-point method (at sentence or intrasentential levels) or the block method (at the text/paragraph level). Finally, the fourth item (*language for comparing*) was reformulated and unified with the parallel item in D1.

Changes to D3: Source-based Writing (SBW)

Finally, some important changes were also made in this dimension. While most raters recommended the inclusion of a dimension focusing on referencing and citation, which is an important aspect of SBW (Pecorari, 2013; Crossley et al., 2021), students in the sample were not instructed to reference or cite the sources, as it is not a common practice in Secondary Education in the context of the study. Regarding major changes, the first item considered the 'use and comprehension of sources', and the importance of doing so accurately and relevantly has been highlighted. More importantly, the second item was divided into two, as the version in V1 considered the use of original ideas as part of the highest level, but as recommended by the majority of the experts, the degree of transformation and the use of original ideas were kept separate. The *degree of transformation* item, therefore, considered lexical and/or structural changes – paraphrasing information in a way relevant to the task (Uludag & McDonough, 2022; Cheong et al., 2021). The new item created, *use of original ideas*, focused on students' ability to integrate their previous knowledge accurately and relevantly to the task.

Using the Rubric: Instructions and Examples

This section presents the final version of the rubric (V3) divided by dimensions and provides examples from the learner corpus gathered for this research. As has been mentioned above, the raters recommended that some important terms, such as *accurate* or *relevant,* be defined to properly use the rubric. As including them in the rubric itself would make it very dense, the scale is accompanied by the following definitions:

(1) Accurate: producing correct language forms following the linguistic norms of the target language, i.e., no errors (Wolfe-Quintero et al., 1998).
(2) Appropriate: the correct use of a linguistic resource according to its function within a context (Nikula et al., 2016).
(3) Complex/simple: related to the sophistication (depth) of the linguistic structures (syntactic and lexical) employed by the writer (Lu, 2017).
(4) Varied: related to the size (diversity) of the linguistic structures (syntactic and lexical) employed by the student (Lu, 2017), i.e., lack of repetition of structures.
(5) Relevant: use of information or ideas that are connected to the main purpose of the text (Leki & Carson, 1997; Macagno, 2016).

Tables 3, 4 and 5 below provide the final version of the rubric for each dimension[3]. To illustrate the rubric performance categories and their pedagogical relevance, three student examples are presented in Table 2. These excerpts are included for illustrative purposes only, as the systematic empirical analysis of the full corpus is reported elsewhere (see Arias-Hermoso et al., 2025a, 2025b). Two examples come from science essays (Ex1 and Ex3) and one from history (Ex2). Ex1 and Ex2 show texts from Year 7 students (12 years old), and Ex3 from a Year 10 student (16 years old).

Focusing on the argumentation dimension, we can observe important differences between the three texts above. While all three texts clearly present the writer's opinion explicitly (3 in D1_ST), their reasoning is very different. Ex1 uses some data to support their claim, but does not elaborate or link to the stance directly (level 1 D1_R), while the other two texts present more complete reasoning structures, with Ex3 including warrants (level 3 in D1_R), such as 'the chocolate cause adiction and this produce obesity, if we don't control it'. None of the texts includes explicit rebuttals, however, they give potential relevant counterarguments (level 2 in D1_CR). Regarding the *language for arguing* item, there is a great difference between Ex1 and Ex2/3. The first example does not use explicit argumentation markers, apart from 'because' twice, whereas the other two texts include more varied and complex lexicogrammar, such as 'in my opinion', 'I don't think that…', 'first of all' or 'apart of their components'. However, Ex2 and Ex3 did not reach the maximum level of this item, as some of the lexicogrammatical resources were not used appropriately or accurately ('apart of', 'becouse').

Regarding comparison skills, we can also see great differences across texts. The first text barely compares the two concepts, while the second and third texts present more clear comparisons between the two concepts by mentioning some differences. However, they are not 'equally elaborated on', a requirement to achieve the highest level in the rubric. Regarding the comparative criteria, three different levels can be seen in the texts, with Ex1 not displaying a clear criterion for the comparisons, Ex2 making comparisons based on implicit criteria (aspects related to *education* and *health*), and Ex3

Table 2

Three Unaltered Examples from the Corpus

Ex 1 (012_E1)	Hello, today I'm compare chicken and chocolate with milk. And wich is healthier between the two foods. The chicken is more healthier because the chocolate with milk haves more calories, more fat and more sugar. The chicken it is used in many healthy diets, it helps the nervous system and the digestive system and it gives energy to the body and the chocolate with milk only it helps to control our cholesterol and it affect our state of mind and because then the chicken is much healthier.
Ex 2 (006_E1)	In my opinion, I prefer to live in todays society, because, now, the economy is based on the service sector, there are many service, and becouse we live more safe. I don't think that people will live better on the middle ages, because, the local work and life were promoted, as well, as the help among neighbours, and, the childrens had to work every day! That is the most sadest thing in the world. And another interesant thing is that only reach people had access to education!! And now, now it is necessary to have studies! Or you can't be nothin, on your life. In the middle ages, people was killed for common illnesses like, flu, diharrea, and more things like that, and now, that doesn't affect us! And common illnesses on our live. Now you live, with 85 years or 90 years old. We are living in a pandemic, an there are more and more and more and… more, new illnesses, live covid 19 and illnesses like that. It was so hard to leave in the midde ages, that wasn't many doctors, and you can die because anithing can kill you.
Ex 3 (098_E4)	I'm going to explain why 100 g of chicken are healthier than 100 g of milk chocolate. First of all, I'm going to compare their components. The milk chocolate has aproximated % 500 more calories than the chicken, the chocolate has 531 kcal and the other only 99,2 kcal. And refering to the total fats the chocolate has literally % 2900 more the proportion is 1 g / 29 g. Looking to other many factors the proportions stay up like in sugar and calcium. But in other factors, The chiekn has more like in proteins and salt. Apart of their components we can see what effects they have in our sistem. The chocolate help us to mantain stable the coresterol and increasing our memory. But the chicken help our nervous sistem and that is very important. One bad thing that have the chicken is that if we eat it raw We can intoxicate with it. The chocolate cause adiction and this produce obesity, if we don't control it. Is true that the chocolate don't envolve animals but yes persons that are exploted in their jobes.

[3] For the complete version of the rubric, access institutional repository: https://hdl.handle.net/20.500.11984/6860

Table 3

D1: Argumentation Skills Rubric, with 0-3 Indicators

D1: ARGUE					
ITEM	**DEFINITION**	**0 - Inexistent, poor**	**1 - Fair**	**2 - Good**	**3 - Excellent**
Stance-taking (D1_ST)	*This is defined as the writer's ability to express their claim or position towards the topic and take a stance.*	The writer's stance towards the topic is not stated, or it is not consistent throughout the text.	The writer's stance towards the topic is difficult to discern, as it is implied.	The writer's stance towards the topic is presented explicitly and is mostly clear.	The writer's stance towards the topic is presented explicitly, clearly and accurately.
Reasoning (D1_R)	*This is defined as the writer's ability to provide supporting data and warrants to justify their stance.*	The writer does not produce data to support their stance or the data is irrelevant, incorrect and/or contradictory.	The writer produces accurate data to support their stance, but it is not elaborated on or not directly linked to their stance.	The writer produces accurate and relevant data to support their stance, and it is elaborated on and/or explicitly linked to their stance.	The writer elaborates on accurate and relevant data to support their stance, including warrants and/or examples for justification purposes.
Counterargumentation and rebuttal (D1_CR)	*This is defined as the writer's ability to acknowledge and refute opposing views to their claim, by providing counterarguments and rebuttals based on data.*	The writer does not mention, acknowledge, or recognise opposing views.	The writer acknowledges that opposing views exist but does not produce counterarguments based on data.	The writer produces accurate and relevant counterarguments, but they are not rebutted or critiqued.	The writer produces accurate and relevant counterarguments, explicitly rebutting and/or critiquing opposing views.
Language for arguing (D1_LA)	*This is defined as the writer's ability to use appropriate, accurate and varied lexicogrammatical resources to produce argumentation elements, including connectives, words to express cause and effect, evaluative language, modalisation...*	The writer does not use lexicogrammatical resources to produce argumentation elements or uses them incorrectly.	The writer uses minimal lexicogrammatical resources to produce argumentation elements, and/or their use might not be completely accurate and/or appropriate.	The writer uses lexicogrammatical resources to produce argumentation elements appropriately and accurately, but they are mostly simple and not varied.	The writer appropriately and accurately uses a variety of simple and complex lexicogrammatical resources to produce argumentation elements.

explicitly stating the criteria (*components* and *effects in our system*). Concerning the comparative structure, Ex1 is somewhat logically structured at the sentence level, but is not completely clear at the paragraph or textual levels. Ex2 and Ex3, in contrast, present a clearer structure, with paragraphs related to a central element and using a point-by-point structure. The structure within the paragraphs, however, does not present comparisons systematically. Finally, students' use of *language for comparing* also differs to a great extent. Ex1 incorrectly uses the comparative form with adjectives (more healthier) once, but uses correct comparative noun forms, 'more calories'. Performing better, Ex2 and Ex3 use more diverse lexicogrammatical resources by including, for example, the words 'prefer', 'looking to other factors', or 'referring to'. They also present some errors, such as 'the most sadest', and, therefore, their use of comparative language is not completely accurate.

The final dimension of the rubric considered SBW skills. Ex1 includes some information from the source, the information is not well used, therefore, introducing certain irrelevant aspects. In contrast, Ex2 and Ex3 present a good amount of information relevant to the task and there are no inaccuracies, even accurately including details from the sources. Regarding the degree of transformation, Ex1 only made minimal linguistic changes, with a great deal of copying. Conversely, Ex 2 and Ex3 both use the information from the source but put it in their own words by introducing some lexical changes. In fact, for instance, the idea that 'The chocolate help us to mantain stable the coresterol and increasing our memory' was originally presented as two ideas: 'It helps us control our cholesterol' and 'it improves our memory'. The student in Ex3 was able to make variations in the wording and introduce the same ideas with their own words. Finally, the first text does not use any original idea not present in the source, while Ex2 and Ex3 do so to a certain extent. Ex2 includes the idea of doctors not being common in the Middle Ages, and Ex3 mentions people being exploited to create chocolate. These ideas are mostly well integrated with the information from the sources, therefore, the texts received a 2.

DISCUSSION

The main objective of the present article has been to report on the validation process of a rubric aimed at assessing three dimensions of ADLs through an expert judgment procedure. Unlike our other empirical studies (Arias-Hermoso et al., 2025a, 2025b) that apply the rubric to corpus data, this article has focused specifically on the validation process, highlighting how expert evaluations can refine and strengthen an educational assessment tool.

Table 4

D2: Comparison Skills Rubric, with 0-3 Indicators

D2: COMPARE					
ITEM	**DEFINITION**	**0 - Inexistent, poor**	**1 - Fair**	**2 - Good**	**3 - Excellent**
Comparative items *(D2_CI)*	*This is defined as the writer's ability to produce comparisons between the two concepts, by including differences and/or similarities, and elaborating on them.*	The writer does not compare the two concepts.	The writer partly compares the two concepts by mentioning some similarities and/or differences.	The writer clearly compares the two concepts by mentioning similarities and/or differences.	The writer clearly compares the two concepts by equally elaborating on several similarities and/or differences among the concepts.
Comparative criteria *(D2_CC)*	*This is defined as the writer's ability to explicitly express the comparative criteria, i.e., the bases of comparison (the criterion according to which the similarities/differences are being made).*	The writer does not produce a clear criterion for comparisons.	The writer produces comparisons based on a criterion, but it is not explicitly stated.	The writer explicitly states the criteria for comparisons.	The writer explicitly states and elaborates on relevant criteria for comparisons.
Comparative structure *(D2_CS)*	*This is defined as the writer's ability to organise the comparisons logically, at sentence, paragraph and text levels.*	The writer does not present structured comparisons throughout the text.	The writer presents comparisons that are somewhat logically structured.	The writer presents point-by-point comparisons that are logically and appropriately structured.	The writer presents point-by-point comparisons that are logically structured at sentence, paragraph and text levels.
Language for comparing *(D2_LC)*	*This is defined as the writer's ability to use appropriate, accurate and varied lexicogrammatical resources to produce comparisons, including compare/contrast connectives, lexical and grammatical forms to express difference/similarity, juxtapositions…*	The writer does not use lexicogrammatical resources to produce comparisons or uses them incorrectly.	The writer uses minimal lexicogrammatical resources to produce comparisons, and/or their use might not be completely accurate and/or appropriate.	The writer uses some lexicogrammatical resources to produce comparisons appropriately and accurately, but they are mostly simple and not varied.	The writer appropriately and accurately uses a variety of simple and complex lexicogrammatical resources to produce comparisons.

Table 5

D3: Source-Based Writing Skills Rubric, with 0-3 Indicators

D3: SBW					
ITEM	**DEFINITION**	**0 - Inexistent, poor**	**1 - Fair**	**2 - Good**	**3 - Excellent**
Use and comprehension of sources *(D3_UC)*	*This is defined as the writer's ability to convey their comprehension of the sources through the use of sources accurately in a relevant way.*	The writer's text presents indicators of serious comprehension problems or does not use the source at all.	The writer's text presents (little) information from sources somewhat relevant to the task, but there are major inaccuracies in conveying information from the sources.	The writer's text presents some information from sources relevant to the task, but there are minor inaccuracies in conveying details from the sources.	The writer's text presents substantial information from sources always relevant to the task, and details are accurately presented.
Degree of transformation *(D3_DT)*	*This is defined as the writer's ability to transform the information from the sources via paraphrasing (lexical and structural changes).*	The writer copies most information directly from the source and makes minimal or no linguistic changes.	The writer copies most information directly from the source, but there is some linguistic change.	The writer integrates information from the source by mostly paraphrasing with lexical or structural changes.	The writer accurately integrates information from the source by transforming it with both lexical and structural changes.
Inclusion of original ideas *(D3_O)*	*This is defined as the writer's ability to include ideas not present in the sources, (original ideas from the writer's background) by appropriately integrating external knowledge into the text.*	The writer does not include external information.	The writer includes some external information, but it is not appropriately integrated with source information.	The writer includes some external information in the text, and it is mostly well integrated with information from the sources.	The writer includes some external information in the text, and it is always appropriately integrated with information from the sources.

The first key finding is related to the validation process per se and how expert judgement proved crucial in developing a validated and multi-dimensional rubric. With the input from 13 leading scholars in the field of ADLs, the iterative process made it possible to identify ambiguous descriptors, particularly in clarity, and to reformulate them into more operational items. As highlighted by previous research (Lázaro Cantabrana & Gisbert Cervera, 2015; Uludag & McDonough, 2022), the design of assessment tools guided by experts' views provides a systematic foundation to create reliable and sound rubrics, a crucial aspect to ensure quality (Landeta & Lertxundi, 2023). Just as comparative judgement studies have been successful in drawing teachers' and researchers' insights to refine ADL-assessment (e.g., Llinares et al., 2024), the present study contributes to the growing body of research demonstrating the usefulness of systematic expert involvement for rubric development. This can be seen in the expert consensus across items, which increased from 18.6 % agreement in V1 to 76.6 % in V2, which represent a crucial improvement in rubric quality and construct validity.

A second important finding concerns the nature of *quality* in each of the three dimensions. As ADL-related skills are often complex and multi-componential (del Pozo, 2024; Nikula et al., 2024), it is crucial that tools include multiple aspects or components that clearly operationalise what constitutes a *great* argument, comparison or use of sources. The inclusion of multiple items underscores this multidimensional nature and improves construct validity (Paltridge & Phakiti, 2015). In particular, experts' consensus converged on the need for complete descriptors with various components. For instance, experts highlighted that high-quality performance in argumentation involves not only presenting clear claims but also justifying evidence and presenting rebuttals. This resonates with previous research on both rubric-based and component-based assessment of ADLs, regarding the integration of complete argument structures (Jiménez-Aleixandre & Erduran, 2007; Polias, 2016; Sendur et al., 2021), a structured account of comparative items and criteria (Gerns, 2023a; Chen & Zhou, 2022), and the role of source use and integration (Chau et al., 2022; Cheong et al., 2021). The inclusion of different items per dimension helps track student performance at various levels or ages to capture more nuanced patterns (see Arias-Hermoso et al., 2025b). In fact, as shown by empirical research on ADL development, different items within the same dimension may develop differently, for example, *reasoning* over *counterargumentation* (Arias-Hermoso et al., 2025a). The rubric, supported by the 13 experts' voices, therefore, emphasises the relevance of multiple components or items to address ADLs, and helps to define both conceptually and operationally the three key dimensions under study.

CONCLUSION

The present study aimed to validate a multidimensional rubric to assess three key dimensions of ADLs, i.e., argumentation skills, comparison skills and source-based writing skills. Through an expert-judgement procedure that engaged 13 scholars in two rounds, we demonstrated that expert feedback is crucial to refine ADL descriptors. The rubrics advance work on previous descriptors or rubrics by confirming and emphasising the multidimensional nature of ADLs, as well as by incorporating SBW into the construct. Methodologically, the study shows that expert-based feedback can produce rubrics that are both theoretically grounded and usable across languages and disciplines. Our study, therefore, combining research from different perspectives and disciplines, provides a tool that can foster the integration of content and language for assessment and teaching purposes.

The study, therefore, has important implications for both theory and practice. At a theoretical level, the rubric confirms the multidimensional nature of ADLs and CDFs. In addition to that, to our knowledge, no study has yet focused on SBW within ADLs, and this paper offers interesting points for their connection that still require further research. At a more practical level, this study offers great opportunities for research and classroom practices. The rubrics validated here are of great benefit for researchers interested in CDFs, ADLs and SBW, and could be used and adapted to different linguistic, educational and disciplinary contexts. The rubrics can be applied in various educational levels and contexts, and can be of great benefit for primary, secondary and higher education teachers, as well as for material designers, curriculum developers and other researchers in academic writing. Furthermore, the rubrics and the accompanying illustrative examples could be beneficial for language and content teachers for classroom instruction and assessment. Based on the rubrics, for example, content (e.g., History) teachers could design lessons focusing on comparing two historical phenomena, instructing them on how to highlight differences and similarities.

Despite the implications, some limitations to this research have to be addressed. One limitation is that it did not include empirical analyses of student performance, as the focus of the paper was on validating the rubric through expert judgement. Another limitation is related to the lack of psychometric modelling (for example, many-facet Rasch measurements) beyond descriptive statistics, which future research could combine with expert judges' qualitative input. An aspect that is both a contribution and a limitation is that the rubrics are not language- and discipline-specific. While this allows for their use in multiple languages and disciplines, they still may not fully capture subject-specific characteristics of CDF production, since what constitutes defining in science might be different from what is expected in history. CDF-based assessments could be adapted to the needs of specific disciplines. This is in line with previously mentioned CEFR-style assessment grids based on CDFs for mathematics, science and history. These grids, however, do not consider various elements within each CDF. Building on this, future studies could bring these perspectives together by building

subject-specific rubrics that consider multiple components for each CDF on top of the two presented in this study. Ideally, future research could try to create similar rubrics for the 7 CDF categories for multiple disciplines, which would also highlight the similarities and differences across disciplines, therefore proving the cross-disciplinary nature of the CDF construct. In addition, we believe that this process could also benefit from not only expert academics' judgement, but also from classroom-based teachers' ratings to create these rubrics.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the 13 experts who generously contributed to the validation of this rubric, for their time, effort and expertise. Thank you for your support and collaboration.

This research received funding from the Basque Government under Grants [IT1664-22 & PRE_2021_1_0001] and by the Spanish Ministry of Science, Innovation and Universities [PID2019-111655RA-I00].

ETHICS STATEMENT

This research received the approval of the Ethics Committee of Mondragon University (ID: IEB20231218). Expert participants (the 13 judges) were informed about the nature of the task and informed consent was given to participate. Student participants from the intended corpus were also informed about the nature of the research, and their parents and/or guardians signed an informed consent for their participation. Data was treated solely for research purposes and anonymity was ensured during the whole process.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTIONS

Roberto Arias-Hermoso: conceptualisation; methodology; validation; analysis; writing (original draft); funding acquisition.

Eneritz Garro Larrañaga: conceptualisation; methodology; validation; analysis; reviewing and editing; funding acquisition.

Ainara Imaz Agirre: conceptualisation; methodology; validation; analysis; writing; funding acquisition.

REFERENCES

Allagui, B. (2019). Investigating the quality of argument structure in first-year university writing. In S. Hidri (Ed.), *English language teaching research in the Middle East and North Africa* (pp. 179–199). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-98533-6_9

Arias-Hermoso, R., Imaz Agirre, A., & Garro Larrañaga, E. (2024). A comparison between input modalities and languages in source-based multilingual argumentative writing. *Assessing Writing, 60,* 100813. https://doi.org/10.1016/j.asw.2024.100813

Arias-Hermoso, R., Garro Larrañaga, E., Imaz Agirre, A., & Dalton-Puffer, C. (2025a). Producing cognitive discourse functions in disciplinary Basque writing: developmental patterns across secondary education and L1 profiles. *International Journal of Bilingual Education and Bilingualism, 28*(7), 744–763. https://doi.org/10.1080/13670050.2025.2488736

Arias-Hermoso, R., Imaz Agirre, A., & Garro Larrañaga, E. (2025b). Multilingual disciplinary literacies: exploring developmental patterns of science writing across secondary education. *Journal of Multilingual and Multicultural Development,* 1–19. https://doi.org/10.1080/01434632.2025.2481200

Bauer-Marschallinger. (2022). *CLIL with a capital I: Using cognitive discourse functions to integrate content and language learning in CLIL history education* [Unpublished doctoral thesis]. University of Vienna.

Breeze, R., & Dafouz, E. (2017). Constructing complex Cognitive Discourse Functions in higher education: An exploratory study of exam answers in Spanish- and English-medium instruction settings. *System, 70*, 81-91. https://doi.org/10.1016/j.system.2017.09.024

Chan, S., & May, L. (2022). Towards more valid scoring criteria for integrated reading-writing and listening-writing summary tasks. *Language Testing, 40*(2), 410-439. https://doi.org/10.1177/02655322221135025

Chau, L.T., Leijten, M., Bernolet, S. & Vangehuchten, L. (2022). Envisioning multilingualism in source-based writing in L1, L2, and L3: The relation between source use and text quality. *Frontiers in Psychology, 13*, 01-20. https://doi.org/10.3389/fpsyg.2022.914125

Chen, M., & Zhou, H. (2022). Comparison-and-contrast in research articles of applied linguistics: A frame-based analysis. *Lingua, 276*, 103387. https://doi.org/10.1016/j.lingua.2022.103387

Cheong, C. M., Zhu, X. & Liao, X. (2018). Differences between the relationship of L1 learners' performance in integrated writing with both independent listening and independent reading cognitive skills. *Reading and Writing, 31,* 779–811. https://doi.org/10.1007/s11145-017-9811-8

Cheong, C. M., Zhu, X. & Xu, W. (2021). Source-based argumentation as a form of sustainable academic skill: An exploratory study comparing secondary school students' L1 and L2 writing. *Sustainability, 132*, 2869. https://doi.org/10.3390/su132212869

Coffin, C. (2009). *Historical discourse: The language of time, cause and evaluation*. Bloomsbury.

Connolly, T. (2019). *Die Förderung vertiefter Lernprozesse durch Sachfachliteralität: Eine vergleichende Studie zum expliziten Scaffolding kognitiver Diskursfunktionen im bilingualen Chemieunterricht am Beispiel des Erklärens* [] [Unpublished doctoral dissertation]. Johannes Gutenberg-Universität Mainz.

Crossley, S. A., Tian, Y., & Wan, Q. (2022). Argumentation features and essay quality: Exploring relationships and incidence counts. *Journal of Writing Research, 14*(1), 1–34. https://doi.org/10.17239/jowr-2022.14.01.01

Dalton-Puffer, C., & Bauer-Marschallinger, S. (2019). Cognitive discourse functions meet historical competences: Towards an integrated pedagogy in CLIL history education. *Journal of Immersion and Content-Based Language Education, 7*(1), 30–60. https://doi.org/10.1075/jicb.17017.dal

Dalton-Puffer, C., Bauer-Marschallinger, S., Brückl-Mackey, K., Hofmann, V., Hopf, J., Kröss, L., & Lechner, L. (2018). Cognitive discourse functions in Austrian CLIL lessons: Towards an empirical validation of the CDF construct. *European Journal of Applied Linguistics, 6*(1), 5–29. https://doi.org/10.1515/eujal-2017-0028

Dalton-Puffer, C., Hüttner, J., & Nikula, T. (2024). The conceptualisation of disciplinary literacies in CLIL. In J. Hüttner & C. Dalton-Puffer (Eds.), *Building disciplinary literacies in content and language integrated learning* (pp. 1-25). Routledge.

Dalton-Puffer, C. (2013). A construct of cognitive discourse functions for conceptualising content-language integration in CLIL and multilingual education. *European Journal of Applied Linguistics, 1*(2)*,* 216–253. https://doi.org/10.1515/eujal-2013-0011

Dalton-Puffer, C. (2016). Cognitive discourse functions: Specifying an integrative interdisciplinary construct. In T. Nikula, E. Dafouz, P. Moore, & U. Smit. (Eds.), *Conceptualising Integration in CLIL and Multilingual Education* (pp. 29–54). Multilingual Matters.

deBoer, M. (2020). Teacher-based assessment of learner-led interactions in CLIL: The power of cognitive discourse functions. In M. deBoer & D. Leontjev (Eds.), *Assessment and learning in content and language integrated learning (CLIL) classrooms* (pp. 203–223). Springer. https://doi.org/10.1007/978-3-030-54128-6_10

del Pozo, E., & Llinares, A. (2021). Assessing students' learning of history content in Spanish CLIL programmes: A content and language integrated perspective, in C. Hemmi, & D. L. Banegas (Eds.), *International Perspectives on CLIL. International perspectives on English language teaching.* Palgrave MacMillan. https://doi.org/10.1007/978-3-030-70095-9_3

del Pozo, E. (2024). Assessment in CLIL: the pending subject in bilingual education? A case study. *Revista De Educación, 1*(403), 231–257. https://doi.org/10.4438/1988-592X-RE-2024-403-605

Evnitskaya, N., & Dalton-Puffer, C. (2024). CLIL learners' categorizations: Writing about history in English across three grade levels in Spanish bilingual schools. In J. Hüttner & C. Dalton-Puffer (Eds.), *Building disciplinary literacies in content and language integrated learning* (pp. 60-79). Routledge.

Evnitskaya, N., & Dalton-Pufrer, C. (2023). Cognitive discourse functions in CLIL classrooms: Eliciting and analysing students' oral categorizations in science and history. *International Journal of Bilingual Education and Bilingualism, 26*(3), 311–330. https://doi.org/10.1080/13670050.2020.1804824

Fang, Z. & Coatoam, S. (2013). Disciplinary literacy: what you want to know about it. *Journal of Adolescent and Adult Literacy, 56* (8), 627-632. https://doi.org/10.1002/JAAL.190

Gerns, P. (2023a). Building scientific knowledge in English: Integrating content, cognition and communication in secondary school CLIL biology. *Journal of Language and Education, 9*(3), 52–78. https://doi.org/10.17323/jle.2023.17569

Gerns, P. (2023b). Qualitative insights and a first evaluation tool for teaching with cognitive discourse function: "Comparing" in the CLIL science classroom. *Porta Linguarum, 40*, 161–179. https://doi.org/10.30827/portalin.vi40.26619

Granados, A., & Lorenzo, F. (2024). A functional description of disciplinary literacy in history: Applications of the Common European Framework of Reference for Languages to content and language integrated learning courses. In J. Hüttner & C. Dalton-Puffer (Eds.), *Building disciplinary literacies in content and language integrated learning* (pp. 83–100). Routledge.

Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge.

Harsch, C., Koval, V., Kanistra, P. V., & Delgado-Osorio, X. (2024). Validating an integrated reading-into-writing scale with trained university students. *Assessing Writing*, *62*, 100894. https://doi.org/10.1016/j.asw.2024.100894

Imaz Agirre, A., Arias-Hermoso, R., & Ipiña, N. (2025). The effect of an intervention focused on academic language on CAF measures in the multilingual writing of secondary students. *International Review of Applied Linguistics in Language Teaching, 63*(2), 1373–1396. https://doi.org/10.1515/iral-2023-0137

Jackson, D. O. (2024). The longitudinal development of argumentative writing in an English for academic purposes course in Japan. *System*, 103482. https://doi.org/10.1016/j.system.2024.103482

Jiménez-Aleixandre, M. P., & Erduran, S. (2007) Argumentation in science education: An overview. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education* (pp. 3–27). Springer.

Kato, M. (2018). Exploring the Transfer Relationship of Summarizing Skills in L1 and L2. *English Language Teaching Archives, 11*(10). https://doi.org/10.5539/elt.v11n10p75

Kuiken, F., & Vedder, I. (2020). Scoring approaches: Scales/rubrics. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (1st ed., pp. 10). Routledge. https://doi.org/10.4324/9781351034784

Landeta, J., & Lertxundi, A. (2024). Quality indicators for Delphi studies. *Futures & Foresight Science, 6*(1), e172. https://doi.org/10.1002/ffo2.172

Lázaro Cantabrana, J. L., & Gisbert Cervera, M. (2015). El desarrollo de la competencia digital docente a partir de una experiencia piloto de formación en alternancia en el Grado de Educación [Developing teachers' digital competence through a pilot alternating training experience in the Bachelor's degree in education]. *EDUCAR, 51*(2), 321-348. http://dx.doi.org/10.5565/rev/educar.725

Lersundi, A. (2022). *Analysis of subject-specific literacies in a multidisciplinary project in upper-secondary education. Case Study.* [Unpublished Doctoral thesis]. Mondragon Unibertsitatea. https://hdl.handle.net/20.500.11984/5964

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. In E. Cano & G. Ion (Eds.), *Innovative practices for higher education assessment and measurement* (pp. 119-138). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-5225-0531-0.ch007

Li, D. (2022). A review of academic literacy research development: From 2002 to 2019. *Asian-Pacific Journal of Second and Foreign Language Education, 7*(1), 1-22. https://doi.org/10.1186/s40862-022-00130-z

Liu, F., & Stapleton, P. (2014). Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test. *System, 45*, 117-128. https://doi.org/10.1016/j.system.2014.05.005

Llinares, A., Morton, T., & Whittaker, R. (2024). Fostering language awareness for integration through teacher-researcher collaboration in a Spanish bilingual education context. *Language Awareness*, 1–21. https://doi.org/10.1080/09658416.2024.2385766

Llinares, A., & Morton, T. (2024). Assessing CLIL students' expression of Explore across languages and school disciplines: An interdisciplinary approach, in J. Hüttner & C. Dalton-Puffer (Eds.), *Building Disciplinary literacies in content and language integrated learning* (pp. 101-118). Routledge. https://doi.org/10.4324/9781003403685

Lorenzo, F., Lorenzo, F., Cvikic, L., Llinares, A., De Boer, M., Adadan, E., Arias-Hermoso, R., Ćaleta, M., Demirkol Orak, S., Evnitskaya, N., Glasnović Gracin, D., Granados, A., Guzmán-Alcón, I., Kasprzak, M., Lehesvuori, S., Miloshevska, L., Özdemir, H., Piacentini, V., del Pozo, E., Roquet, H., Bagalová, D., and Ting, T. (2024). *Assessing disciplinary literacy with CEFR descriptors: History, Mathematics and Science. A paper by CLILNetLE Working Group 2*. V1. PHAIDRA repository, University of Vienna. https://doi.org/10.25365/phaidra.532_30

Lorenzo, F. (2017). Historical literacy in bilingual settings: Cognitive academic language in CLIL history narratives. *Linguistics and Education, 37*, 32-41. https://doi.org/10.1016/j.linged.2016.11.002

Lorenzo, F. (2023). Academic language as linguistic capital – A window to social justice: A commentary on "Midadolescents' language learning at school: Toward more just and scientifically rigorous practices in research and education." *Language Learning, 73*(1), 136-143. https://doi.org/10.1111/lang.12560

Meneses, A., Montenegro, M., Acevedo, D., Figueroa, J., & Hugo, E. (2023). Cross-disciplinary language changes in 4th graders as a predictor of the quality of written scientific explanation. *Journal of Writing Research, 15*(1), 105–132. https://doi.org/10.17239/jowr-2023.15.01.05

Moje, E. B. (2015). Doing and teaching disciplinary literacy with adolescent learners: A social and cultural enterprise. *Harvard Educational Review, 85*(2), 254–278. https://doi.org/10.17763/0017-8055.85.2.254

Morton, T., & Nashaat-Sobhy, n. (2024). exploring bases of achievement in content and language integrated assessment in a bilingual education program. *TESOL Quarterly, 58*(1), 5-31. https://doi.org/10.1002/tesq.3207

Morton, T. (2020). Cognitive discourse functions: A bridge between content, literacy and language for teaching and assessment in CLIL. *CLIL Journal of Innovation and Research in Plurilingual and Pluricultural Education, 3*(1), 7-17. https://doi.org/10.5565/rev/clil.33

Morton, T. (2022). Using cognitive discourse functions and comparative judgement to build teachers' knowledge of content and language integration for assessment in a bilingual education program. *Journal of Immersion and Content-Based Language Education, 10*(2), 302-322. https://doi.org/10.1075/jicb.21017.mor

Nashaat-Sobhy, N., & Llinares, A. (2023). CLIL students' definitions of historical terms. *International Journal of Bilingual Education and Bilingualism, 9*, 1–14. https://doi.org/10.1080/13670050.2020.1798868

Nikula, T., Nashaat-Sobhy, N., Minardi, S., Gülle, T., Yalçın, S., Duman, S.K., Bozbıyık, M., Delibegović Džanić, N., Ellison, M., Gerns, P., Gómez, E., Hanušová, S., Kováčiková, E., Kääntä, L., Lin, A., Llinares, A.; Yi Lo, Y., Lopriore, L., Meyer, O., Morton,T., Neville, C., Rannu, K., Sağlamel, H., Sula, G., Sulkunen, S., Pavičić Takač, V., Tiermas, A., Ting, T., Tsagari, D., Yüksel, G., & Žmavc, J. (2024). *Towards an initial operationalisation of disciplinary literacies: A paper by CLILNetLE Working Group 1*. PHAIDRA repository, University of Vienna. https://hdl.handle.net/11353/10.2050621

Paltridge, B., & Phakiti, A. (2015). *Research Methods in Applied Linguistics: A practical resource*. Bloomsbury.

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(3), 217-230. https://doi.org/10.1016/j.jslw.2013.02.003

Polias, J. (2016). *Apprenticing students into science: doing, talking & writing scientifically*. Lexis Education.

Pylonitis, C., & Meyer, O. (2024). Arguing for global citizenship: Mapping deeper learning in the language-as-discipline classroom. In S. Greco & L. Cinganotto (Eds.), *Innovation in education for deeper learning* (pp. 55–78). INDIRE-IUL Press.

Qin, J., & Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System, 38* (3), 444-456. https://doi.org/10.1016/j.system.2010.06.012

Rieder-Marschallinger, S. (2024). Cognitive discourse functions in history CLIL education: Insights from a design-based research study on conceptual links. In J. Hüttner & C. Dalton-Puffer (Eds.), *Building disciplinary literacies in content and language integrated learning* (pp. 195–215). Routledge.

Roca de Larios, J., Coyle, Y., & Garcia, V. (2023). The effects of using cognitive discourse functions to instruct 4th-year children on report writing in a CLIL science class. *Studies in Second Language Learning and Teaching, 12*(4), 597–622. https://doi.org/10.14746/ssllt.2022.12.4.4

Sampson, V., & Clark, D. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*(3), 447–472. https://doi.org/10.1002/sce.20276

Schleppegrell, M. (2004). *The language of schooling: a functional linguistic perspective*. Mahwah: Erlbaum.

Sendur, K. A., van Drie, J., van Boxtel, C., & Kan, K. J. (2020). Historical reasoning in an undergraduate CLIL course: Students' progression and the role of language proficiency. *International Journal of Bilingual Education and Bilingualism, 25*(6), 2058–2074. https://doi.org/10.1080/13670050.2020.1844136

Shanahan, T. & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review, 78*(1), 40-59. https://psycnet.apa.org/doi/10.17763/haer.78.1.v62444321p602101

Smith, S. (2019). *Academic writing genres: Essays, reports & other genres*. Evident Press.

Stapleton, P., & Wu, Y. (2015). Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes, 17*, 12-23. https://doi.org/10.1016/j.jeap.2014.11.006

Steiss, J., Wang, J., Kim, S. G., & Olson, C. B. (2024). U.S. Secondary Students' Source-Based Argument Writing in History. *Written Communication*. https://doi.org/10.1177/07410883241263549

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Trimble, L. (1985). *English for science and technology: A discourse approach*. Cambridge University Press.

Uludag, P., & McDonough, K. (2022). Validating a rubric for assessing integrated writing in an EAP context. *Assessing Writing, 52*, 100609. https://doi.org/10.1016/j.asw.2022.100609

van Weijen, D., Rijlaarsdam, G. & van den Bergh, H. (2019) Source use and argumentation behavior in L1 and L2 writing: A within-writer comparison. *Reading and Writing, 32,* 1635–1655. https://doi.org/10.1007/s11145-018-9842-9

Vercellotti, M. L., & McCormick, D. E. (2021). Constructing analytic rubrics for assessing open-ended tasks in the language classroom. *TESL-EJ, 24*(4).

Whittaker, R., & McCabe, A. (2023). Expressing evaluation across disciplines in primary and secondary CLIL writing: A longitudinal study. *International Journal of Bilingual Education and Bilingualism, 5*(2), 1–18. https://doi.org/10.1080/13670050.2020.1798869