tesol QUARTERLY

Exploring Bases of Achievement in Content and Language Integrated Assessment in a Bilingual Education Program

TOM MORTON
Universidad Autónoma de Madrid
Madrid, Spain

NASHWA NASHAAT-SOBHY
Universitat Politècnica de València
Gandía, Spain

Abstract

This study explores the bases of achievement invoked by teachers when assessing students' work in the context of a bilingual education program where academic subjects are taught through English as a foreign language. During a professional development seminar, teachers judged samples of students' writing in response to tasks that elicited the three cognitive discourse functions (CDFs) of define, evaluate, and explore. The teachers' discourse was analyzed using specialization, a dimension of Legitimation Code Theory-LCT (Maton, Knowledge and Knowers. Towards a realist sociology of education, 2014), a sociological framework for analyzing knowledge practices. Specialization codes provide insight into epistemic relations (knowledge) and social relations (knowers) in educational practices. The results show that within epistemic relations, there was a balance between content and language as bases of achievement. Content quality was emphasized over quantity, language form was emphasized over function, and teachers gave different weights to language depending on the quality of the content. Social relations were also invoked, though less often than epistemic relations. The results suggest that teachers' positioning of students in terms of epistemic and social relations in their assessment practices may have consequences for the equitable treatment of learners in bilingual programs.

doi: 10.1002/tesq.3207

INTRODUCTION

Content and language integrated learning (CLIL) is an approach to bilingual education in which the academic content is taught alongside a second/foreign/additional language, with the objective being the learning and development of both (Morton & Llinares, 2017). Interest in CLIL at the levels of policy, practice, and research has been growing rapidly worldwide over the last two decades and is likely to continue (Dalton-Puffer, Hüttner, & Llinares, 2022). However, as CLIL (especially with English as the language of instruction) spreads across the world, the implications for social equality are just beginning to be explored (Codó, 2022; Hidalgo-McCabe & Fernández-González, 2019; Llinares & Evnitskaya, 2021; Pérez Cañado, 2020). One issue in CLIL practice that has a direct influence on equality is that of assessment. Until recently, assessment has been underresearched in CLIL and has been seen as something of a "blind spot" for practitioners and researchers (Massler, Stotz, & Queisser, 2014, p. 138; Lo & Fung, 2020). At the level of practice, assessment objectives and guidelines for language development are found to be wanting, leading teachers to apply their own bases of achievement (e.g., Otto & Estrada, 2019). In order to improve assessment practice in CLIL, it is necessary to gain a firm understanding of teachers' existing orientations by creating opportunities for them to articulate the bases of achievement underlying the criteria they apply in assessing students' work. In this study, a sociological framework for the exploration and improvement of all types of knowledge practices, Legitimation Code Theory (Maton, 2014) is used to uncover the bases of achievement invoked by a group of teachers working in the context of a bilingual education program in Spain, in which English is used in primary and secondary schools for the teaching of academic subjects such as art, history, and science.

Assessment in CLIL

As far back as 2010, Mohan, Leung, and Slater pointed out that there was a lack of knowledge of adequate theory, analysis, and practice for the integrated assessment of language and content (Mohan, Leung, & Slater, 2010, p. 220). However, since then, the research effort to gain a clearer understanding of the linguistic demands of assessing academic content in an L2 has been gathering pace, for example, in the context of international exams (Shaw & Imam, 2013), assessment instruments for primary CLIL (Massler et al., 2014), and a framework for evaluating content and language demands of assessment tasks as students progress through secondary education (Lo & Fung, 2020; Lo & Lin, 2014). Overall, progress is being made in understanding the mediating effect of linguistic demands on students' performance in expressing content knowledge (Lo, Fung, & Qiu, 2021).

In spite of this progress, there is evidence that CLIL teachers lack conceptual understanding of how to integrate content and language in assessment as well as the adequate materials to do so (Bauer-Marschallinger, 2022). When CLIL teachers do not have access to a theoretical base for the integration of content and language in assessment, they may rely on individual experience and common sense and will therefore lack a common set of criteria that could serve as bases of achievement. This has obvious implications for validity and reliability in assessing students' learning. There may be construct-irrelevant variance (Avenia-Tapper & Llosa, 2015) when content assessment outcomes are affected by teachers using language performance as a criterion for assessment without necessarily being aware of doing so, as when teachers refer to aspects of language performance, such as accuracy or fluency, derived from the assessment tradition of learners in foreign language teaching (Otto & Estrada, 2019). In these cases, language becomes an "invisible" component of assessment (Hönig, 2010).

In terms of consequential validity, decisions based on such components are inherently unfair as some students may be unjustifiably rewarded for accomplished language performance, whereas others may be mistakenly judged as lacking content knowledge or skill due to irrelevant aspects of language performance. Given that student performance involves aspects such as strategic competence and the mastery of general language skills, which are neither academic language nor content knowledge (Massler et al., 2014), it is important that these are not unwittingly used as content assessment criteria. On the other hand, it is important to avoid reducing the language demand so much that it threatens the integrity of the content knowledge being assessed (Shaw & Imam, 2013). Rather than focusing on general language skills, there are strong arguments that the choice of language as an explicit focus of assessment should be directly related to subject-specific thematic patterns and lexis (He & Lin, 2019), and the communicative learning tasks for which specific linguistic features are functional (Avenia-Tapper & Llosa, 2015; Chadwick, 2012; Otto, 2018). Such an approach to integration would be what Leung and Morton (2016) describe as a "higher disciplinary orientation to language/more visible language pedagogy," in which there would be explicit attention to aspects of subject-specific literacy.

A different, but related, issue, is when criteria for assessing learning in CLIL programs are based on those governing performance in the first language (Otto & Estrada, 2019). While this may help to ensure greater reliability in assessment, it also presents a validity problem, as it disregards the fact that the language of instruction is an L2 for the learners and thus presents greater linguistic and cognitive challenges for them (Lo & Fung, 2020). Thus, either focusing on general foreign language learning criteria such as fluency or accuracy, or criteria relevant only to L1 educational contexts, can buy reliability at the expense of validity. Teachers familiarized with these criteria through the use of instruments such as rubrics may consistently apply them to learners' performances, but they may underestimate students' content knowledge either through taking into account construct-irrelevant language criteria or failing to find a balance between cognitive and language demand in assessment tasks.

This can clearly have negative consequences for fairness and equity. At the level of individual students, they may be unfairly labeled as successful or unsuccessful in the subject and be awarded grades that are a poor reflection of their actual content knowledge and may be assigned to inappropriate ability groups or streams. At the societal level, such assessment practices may exacerbate the problem identified by research in some bilingual education contexts, in which some students are systematically capitalized or decapitalized (Martín-Rojo, 2013). Those students who bring with them linguistic capital (e.g., exposure to English in an extra-curricular activity, opportunities to travel, family members who speak English) will be advantaged ("capitalized") over those who do not have access to these extra-curricular linguistic experiences (who will be "decapitalized"). Thus, the lack of theoretical models and practical frameworks for integrating content and language in assessment can potentially have real negative effects on social equity, especially when teachers invoke bases of achievement which are extraneous to the assessment of content knowledge and skills, and which may have been achieved elsewhere by those learners fortunate enough to have had such opportunities.

It is for these reasons that it is important for CLIL teachers to have the opportunity to develop the type of assessment literacy that is specific to content and language integration. Assessment literacy is the "interrelated set of knowledge, skills, and dispositions that a teacher can use to design and implement a coherent and appropriate approach to assessment within the classroom context and the school system" (Pastore & Andrade, 2019, pp. 134–35). For CLIL teachers, this entails having a conceptual model or framework that can bridge content learning and language objectives, and which can be used for the design of instructional and assessment activities. It is only when groups of teachers in CLIL programs share such a conceptual framework that alignment of instructional and assessment activities can take place (Lo, Lui, & Wong, 2019), appropriate scaffolding of language relevant to communicative learning tasks (Lo & Fung, 2020) can be provided, and valid assessment criteria can be consistently applied. Such frameworks for CLIL assessment have been proposed: for example, Lo and Lin's framework combines three levels of cognitive complexity (recall, apply, and analyze) with three levels of linguistic demand (word, sentence, and text) to create a 3 × 3 matrix which allows teachers to analyze the cognitive and linguistic demands of CLIL assessments, gauging these demands as students progress through the curriculum (Lo & Fung, 2020). Coyle and Meyer's (2021) "pluriliteracies" approach also provides an explicit framework for the integrated planning, progression, and assessment of students' content, language, and literacy skills in bilingual programs. DeBoer and Leontjev (2020) propose a conceptualization of classroom assessment in CLIL that combines a classroom-based assessment cycle with the integration matrix proposed by Leung and Morton (2016).

The construct of c*ognitive discourse function* (Dalton-Puffer, 2013) is another framework that allows content and language teachers in a bilingual education program to develop a "common language" from which explicit criteria could be developed for the integrated assessment of content and language. Cognitive discourse functions (CDFs) are the verbal analogs of common content learning objectives usually expressed as verbs denoting cognitive operations. Dalton-Puffer (2013) identified seven "families" of CDF (categorize, define, describe, evaluate, explain, explore, and report), all of which can be described in terms of the learning objectives they represent and the linguistic resources needed for their expression. For example, defining can be seen as a content-based learning objective (providing an accurate definition of a key disciplinary concept) but also in terms of the linguistic resources needed to produce an acceptable definition (e.g., nouns for phenomenon defined and class it belongs to, and grammatical structures such as relative clauses to add extra information). Because they link language functions to learning objectives and can be seen as components of larger text types or genres (Coyle & Meyer, 2021), CDFs can serve as a "bridge" between content-based learning objectives, language, and literacy (Morton, 2020).

The aim of the present study is to explore the underlying principles of the bases of achievement invoked by a group of content and language teachers in discussions after they had assessed samples of primary and secondary students' written work based on tasks that prompted them to produce the CDFs of *define*, *evaluate*, and *explore*. While the professional development activity itself aimed to help develop the teachers' assessment literacy for CLIL, potentially affecting their future assessment practices, the reported empirical study explores the teachers' existing orientations at the time of the intervention. We start from the assumption that, in order to enhance practice in any area of education, we need to have a deep appreciation of the underlying organizing principles of existing practices, which are likely to be prevalent in the type of program under study. In order to reveal the organizing principles underlying the bases of achievement invoked by the teachers in the study, we use the conceptual toolkit of Legitimation Code Theory, a sociological framework for exploring and enhancing all kinds of knowledge practices.

Legitimation Code Theory: Specialization

As explained in Maton (2014), Legitimation Code Theory (LCT) is a response to "knowledge blindness" in education research and practice, a state of affairs in which there is often a greater emphasis on attributes of knowers, as in constructivist or "student-centered" approaches, with the nature of knowledge itself, the forms it takes, and its effects, not being taken into account (Maton, 2014, p. 2). LCT provides a conceptual toolkit that allows researchers to explore the organizing principles that underlie practices and participants' dispositions in different fields of activity, including education, in which knowledge building takes place. LCT provides a powerful set of tools for not only researching but also changing knowledge practices. Although LCT allows for the exploration of knowledge practices along four different dimensions (Specialization, Semantics, Autonomy, and Temporality), the current study draws on the dimension of Specialization.

Specialization allows practices to be explored both in terms of knowers, that is, different kinds of people and their social attributes who may be positioned as more or less legitimate participants in a practice, and knowledge, the specialized concepts, principles, or skills constituting a practice. The organizing principles of practice are revealed in *specialization codes*, which comprise *epistemic relations* (ER) and *social relations* (SR). Epistemic relations refer to the object or focus of practices – the portion of reality to which they are oriented. Social relations are about practices and those who enact them – the subject, the author, or the actor. Both types of relations can vary in strength so if there is a strong focus on the object of practice (what is to be known), it can be labeled as ER+. A weaker focus on epistemic relations is labeled ER−. Likewise, a strong emphasis on social relations is labeled SR+ and a weaker emphasis is SR−. As practices can vary

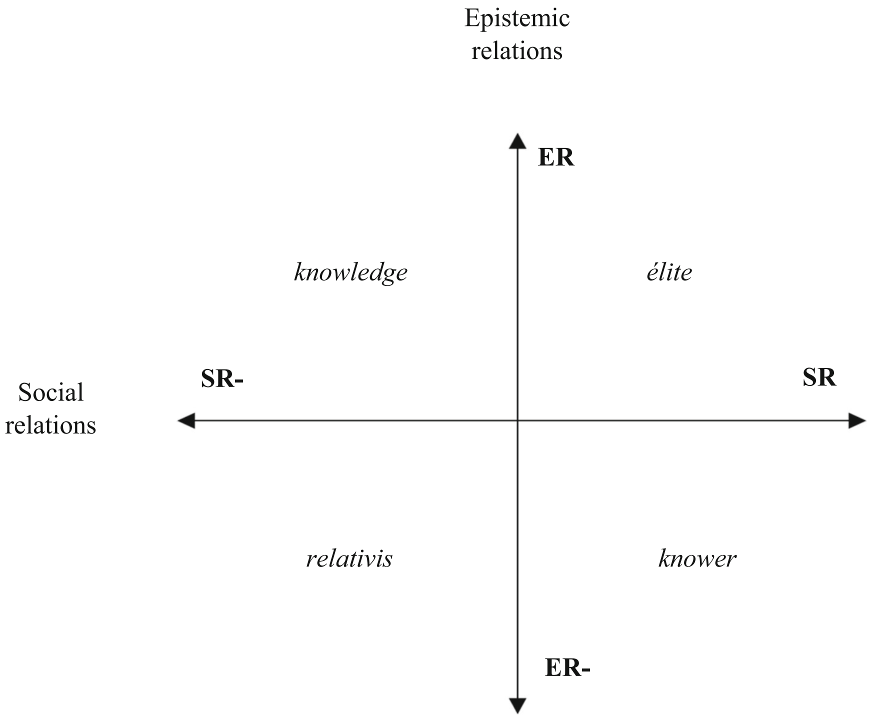


FIGURE 1. Specialization codes (Maton, 2014, p. 30).

across both continua at the same time, this yields four specialization codes (Figure 1).

Knowledge codes (ER+, SR−) emphasize specialized knowledge of objects of study and downplay attributes of the actors involved. What is to be known is important, who you are, less so. Knower codes (ER−, SR+) downplay specialized knowledge as a basis of achievement and instead place the emphasis on attributes of the actors involved, whether these are seen as born (natural talent or gift), cultivated (as in having acquired a certain taste), or social (e.g., gender). Élite codes (ER+, SR+) emphasize both specialized knowledge of an object of study and having the right kind of attributes as a knower. You need to know your stuff and be the right kind of person. Relativist codes (ER−, SR−) downplay both specialized knowledge and knower attributes. What is to be known is not seen as important nor is being a particular kind of knower.

The Specialization dimension of LCT offers a powerful set of tools for revealing the organizing principles underlying the bases of achievement when teachers assess examples of students' work. It can be argued that the conundrums and dilemmas faced by CLIL teachers in integrating content and language for assessment are related to their dispositions concerning the types of learners they deal with and the types of knowledge involved. The “blind spot” in CLIL assessment may reflect deeper knowledge blindness in that the complex relationships between different types of knowledge and knowers are only dimly seen, and need to be revealed with greater clarity if practices are to be improved and negative consequences for social equity avoided. Looking at CLIL and language teachers’ assessment practices through the lens of epistemic and social relations can disentangle some of the complex connections between the different types of knowledge (content and language) invoked as bases of achievement, and how students are positioned as types of knowers in relation to this knowledge. With these key ideas in mind, the study was guided by the following research questions:

In the context of a bilingual education program, in which academic subjects are taught in English:

1. What is the overall balance between content and language knowledge as bases of achievement in the teachers’ accounts of their assessment decisions?
2. What bases of achievement do teachers invoke in terms of students’ content knowledge?
3. What bases of achievement do teachers invoke in terms of students’ language knowledge?
4. What social relations (if any) do teachers invoke as bases of achievement when assessing students’ work?

METHODOLOGY

The data analyzed for the study were collected in the context of a professional development intervention for teachers in the bilingual education program administered by the regional government of Madrid (Spain). The intervention consisted of six 2-h workshops held monthly between November 2019 and April 2020 (the final session was held online due to the Covid-19 pandemic). The participants were seven teachers from schools that offered bilingual education and a team of researchers from a university department. The teachers, all female, were from both primary and secondary schools and taught a range of language and nonlanguage subjects (Table 1).

The goal of the series of workshops was to create a space for integrated content and language collaboration focused on assessment

TABLE 1
Teacher Participants in the Study

Teacher	Subject(s)	Educational level	Experience teaching content in English
A	Science (English medium)	Primary	11 years
B	Science (English medium)	Primary	19 years
C	Spanish language arts	Secondary	Not applicable
D	Art (English medium)	Secondary	4 years
E	History (English medium)	Secondary	15 years
F	English language	Secondary	Not applicable
G	English language	Secondary	Not applicable

among teachers (language and nonlanguage) in bilingual schools. The main conceptual tool used for promoting a more integrated approach to assessment was the construct of cognitive discourse functions as described above. At the beginning of the seminar (workshops 1 and 2), the teachers reflected individually and in groups on the criteria of success for answers that they provided to three prompts for the three CDFs from three subject areas (art, history, and science). Their reflections and discussions were guided by three questions: (1) What information/ideas must a response "contain" to be complete? (2) What criteria would you apply to evaluate the linguistic aspects of the answer? (3) What other criteria do you have? This step was necessary to encourage them to reflect and familiarize themselves with the task ahead in the judging sessions.

In order to provide the participating teachers with the opportunity to articulate the criteria they apply in assessing students' written work, we used the technique of comparative judgment. Comparative judgment is a process in which judges are presented with pairs of responses to a task (usually student scripts) and are asked to decide which is better. Following repeated comparisons, the resulting data are statistically modeled, and responses are ranked on a scale of relative quality. It is claimed that comparative judgment is a more reliable method of assessing students' work than traditional absolute judgment, such as using rubrics and marking scales (Wheadon, de Moira, & Christodoulou, 2020). The teachers used software available at the website of the company No More Marking Ltd https://www.nomoremarking.com/ to judge samples of students' writing in response to tasks that had elicited the CDFs of *define*, *evaluate*, and *explore*. After each judging exercise, the resulting rankings were shown to the teachers in the workshop sessions, and they were asked to discuss reasons why the students' work was ranked in this way. For example, grade 6 students were asked to define an ecosystem, and their

TABLE 2
Teachers' Judgments of Students' Definitions of "Ecosystem" Ranked

A.	An ecosystem is like a food change, in an ecosystem we have producers, consumers, and decomposers	0
B.	An ecosystem is all the living things (plants and animals) in a give area	1,1
C.	An ecosystem is a complex set of relationships among the living resources, habitats, and residents	2,2
D.	An ecosystem is a community formed of a habitat, living things, and the interacting between the different living things themselves and the habitat	3,8
E.	An ecosystem is some space that has some water, forest, plants, air, living things, and rocks	5,6
F.	An ecosystem includes all the living things in a given area, interaction with each other, and also with their nonliving things environment	6,2
G.	An ecosystem is an area with lots of living things interacting with each other. It also includes nonliving environments such as the weather, the climate, ...	10

responses were ranked by the teachers as seen in Table 2, with example G receiving the highest rating.

This judging process allowed for the articulation of criteria underlying judgment decisions in open discussion, potentially providing the means to access the underlying principles of the teachers' knowledge practices in terms of the bases of achievement they oriented to in assessing the students' work. The focus was not on the criteria invoked by individual teachers, nor was the intention to see any teachers as representative of any particular group but the aim was rather to identify patterns emerging in the articulation of their practices. The analysis centered on the meanings being exchanged in the discourse of three of the sessions, in which the teachers discussed their rankings of the students' productions of the CDFs of *define*, *evaluate*, and *explore*, respectively. These discussions were audio recorded and the transcriptions are the corpus analyzed in this study, which is summarized in Table 3.

TABLE 3
The Corpus of the Three Sessions

Session (CDF)	Time	Number of words
1. Define	37 min	4,886
2. Evaluate	83 min	9,687
3. Explore	92 min	10,065
Total	212 min	24,638

Data Analysis

The transcripts of the three sessions were uploaded to the text annotation software program UAM-Corpus Tool (O'Donnell, 2021). This software allows researchers to create their own coding schemes with which to analyze (usually) small corpora collected for specific research projects. The unit of analysis was teachers' articulated opinions as to what constitutes bases of achievement in relation to:

- individual examples (e.g., G is good/better because ...)
- groupings of examples (e.g., G and F are better than C because ...)
- general statements of what constitutes quality not directly related to specific examples.

Three main coding categories emerged upon initial inspection: content, language, and student attributes. We then added two further subcategories to each of these in the second round of coding (content quality and quantity; language form and functions; and individual and societal factors) and refined the aspects related to them when needed, represented in the first three left columns in Table 4. Where both content and language were articulated together as bases of achievement, we used double coding. When the teachers' comments were merely (dis)agreeing with a previous comment from a colleague or a researcher, these were excluded from the coding, unless they elaborated on the reasons for their view by articulating additional or different success criteria. All researchers' comments were excluded, as our focus was on the bases of achievement as articulated by the teachers. The researchers' role in these sessions was to let the teachers take the lead in discussing their own criteria, and they restricted their contributions to mostly prompts to move things on if the discussion flagged, to show interest in certain contributions, and to keep the discussion on track, moving from one set of responses to another.

As Maton and Chen (2016) show, it is not advisable to apply LCT concepts directly to data. First, the data need to be allowed to "speak," using categories that are closer to the context and concerns of the

TABLE 4

Basis of Achievement Invoked in the Teachers' Reflections Across Content and Language

	N = 207	%
Content	112	54
Language	95	46

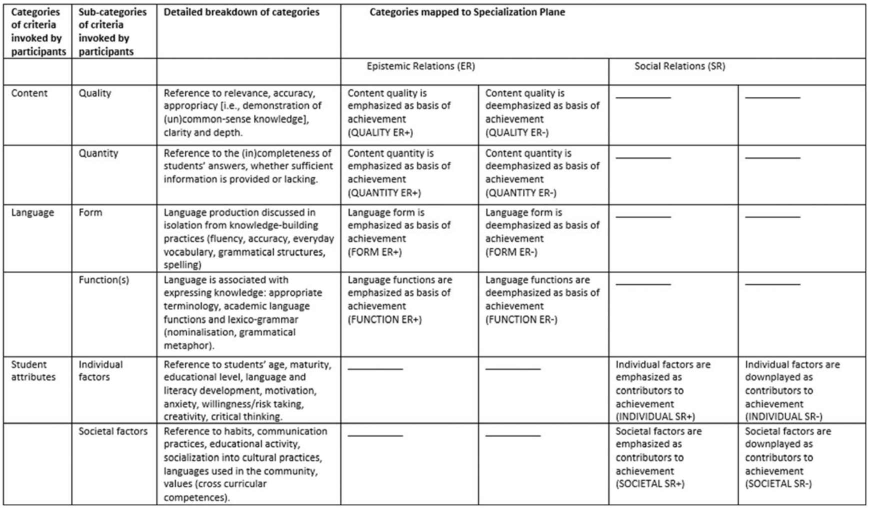
practices being investigated. For this reason, it was important to establish the coding scheme described above. We then distinguished between epistemic relations (ER) and social relations (SR) by coding the focus invoked by the teacher. For example, when a teacher said: “I expect the students to use ‘fewer’ in C because they have studied that ‘less’ is for the uncountable,” the segment is coded “SR” as the emphasis is not on the use of grammar but on the educational practice in which students have been engaged. In contrast, a reflection such as “Student B used accurately the second conditional: If people had more plots, land . . . , they were Patricians,” the segment is coded as “ER” as the emphasis is on the student’s ability to apply knowledge of a grammatical rule in writing. In this sense, Table 5 is what in LCT terms is called a “translation device,” a way to mediate between the theoretical framework of LCT and a corpus of real data, from which its own categories emerge.

RESULTS

In this section, we first present the overall balance between content and language as bases of achievement in the teachers’ reflections, followed by the specific aspects they invoked in each of these two criteria as well as in that of student attributes.

TABLE 5

Translation Device: Coding categories and subcategories mapped against LCT Specialization codes

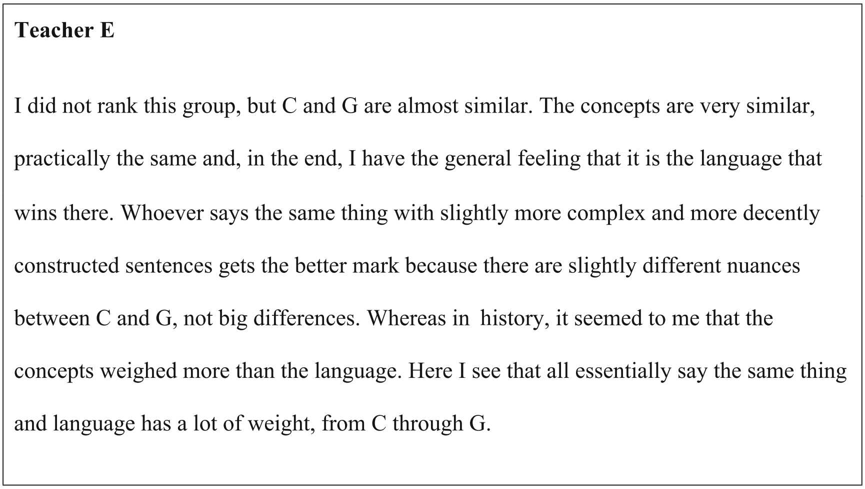


Categories of criteria invoked by participants	Sub-categories of criteria invoked by participants	Detailed breakdown of categories	Categories mapped to Specialization Plane			
			Epistemic Relations (ER)		Social Relations (SR)	
Content	Quality	Reference to relevance, accuracy, appropriacy [i.e., demonstration of (un)common-sense knowledge], clarity and depth.	Content quality is emphasized as basis of achievement (QUALITY ER+)	Content quality is deemphasized as basis of achievement (QUALITY ER-)	—	—
	Quantity	Reference to the (in)completeness of students’ answers, whether sufficient information is provided or lacking.	Content quantity is emphasized as basis of achievement (QUANTITY ER+)	Content quantity is deemphasized as basis of achievement (QUANTITY ER-)	—	—
Language	Form	Language production discussed in isolation from knowledge-building practices (fluency, accuracy, everyday vocabulary, grammatical structures, spelling)	Language form is emphasized as basis of achievement (FORM ER+)	Language form is deemphasized as basis of achievement (FORM ER-)	—	—
	Function(s)	Language is associated with expressing knowledge: appropriate terminology, academic language functions and lexico-grammar (nominalisation, grammatical metaphor).	Language functions are emphasized as basis of achievement (FUNCTION ER+)	Language functions are deemphasized as basis of achievement (FUNCTION ER-)	—	—
Student attributes	Individual factors	Reference to students’ age, maturity, educational level, language and literacy development, motivation, anxiety, willingness/risk taking, creativity, critical thinking.	—	—	Individual factors are emphasized as contributors to achievement (INDIVIDUAL SR+)	Individual factors are downplayed as contributors to achievement (INDIVIDUAL SR-)
	Societal factors	Reference to habits, communication practices, educational activity, socialization into cultural practices, languages used in the community, values (cross curricular competences).	—	—	Societal factors are emphasized as contributors to achievement (SOCIETAL SR+)	Societal factors are downplayed as contributors to achievement (SOCIETAL SR-)

The Balance between Content and Language Knowledge as Bases of Achievement

We coded a total of 250 instances of teachers invoking bases of achievement, 207 of which were almost equally divided among content (112) and language (95). Given that most of the participants were content teachers (5 of the 7 teachers), it was slightly surprising to find such a strong emphasis on language (Table 4).

The interplay between content and language as bases of achievement can be seen in the extracts below. In Extract 1, a history teacher conveys her thoughts about two students' performances (C and G). She observes that though content knowledge (QUALITY ER+) in both C and G is similar, G is ranked higher because it has a more complex grammatical structure (FORM ER+).



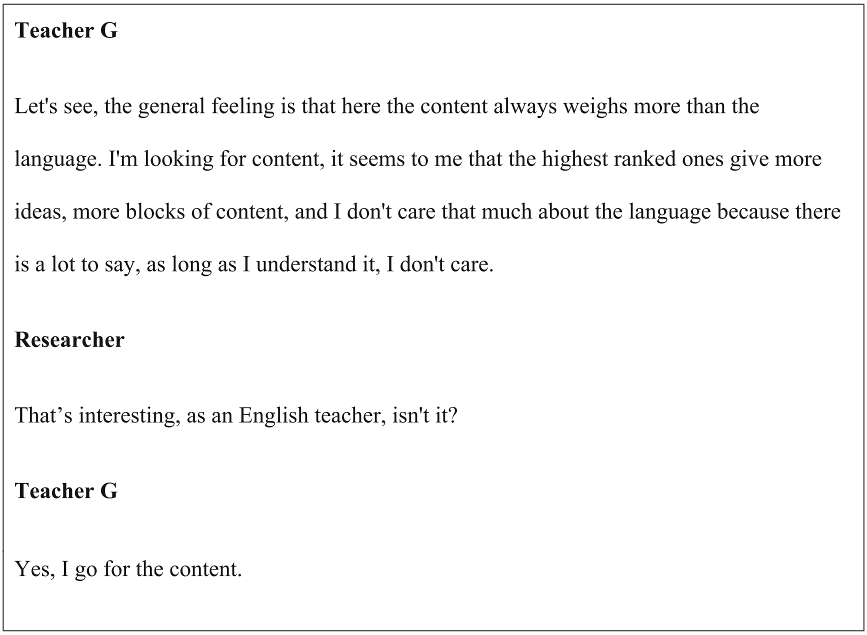
Teacher E

I did not rank this group, but C and G are almost similar. The concepts are very similar, practically the same and, in the end, I have the general feeling that it is the language that wins there. Whoever says the same thing with slightly more complex and more decently constructed sentences gets the better mark because there are slightly different nuances between C and G, not big differences. Whereas in history, it seemed to me that the concepts weighed more than the language. Here I see that all essentially say the same thing and language has a lot of weight, from C through G.

Extract 1.

In the same vein, commenting on a group of performances in the same session (Extract 2), an English teacher observes that she tends to focus more on content when students provide more ideas that provide subject-specific information (QUALITY ER+), and she downplays the focus on language forms (FORM ER−) as long as the students communicate the intended message clearly.

One of the teachers commented that when evaluating students' performance there is "a disconnection between content and language." Rather than a "disconnection," we see that the teachers give different



Teacher G

Let's see, the general feeling is that here the content always weighs more than the language. I'm looking for content, it seems to me that the highest ranked ones give more ideas, more blocks of content, and I don't care that much about the language because there is a lot to say, as long as I understand it, I don't care.

Researcher

That's interesting, as an English teacher, isn't it?

Teacher G

Yes, I go for the content.

Extract 2.

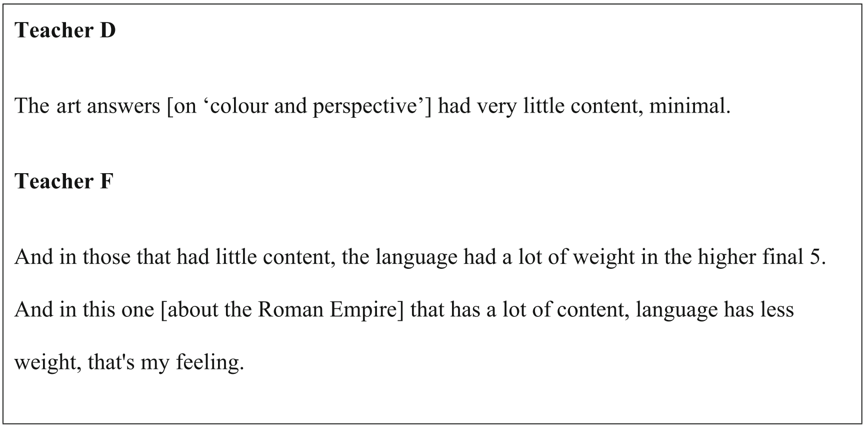
weights to language depending on the quantity and quality of the content. When the content is substantial, they downplay language as a basis of achievement, but when the content is weak, they give more weight to it (Extracts 1–3).

In relation to weighing content and language, the teachers additionally bring up the variation in students' overall performance across two subjects (art and history), thus leading them to focus more on language in art and more on content in history.

Bases of Achievement for Content Knowledge

As shown in Table 6, most of the teachers' reflections on content performance centered on quality (83.93%), not quantity (16.07%), and both unsurprisingly emphasized (ER+) as necessary bases of achievement in the majority of the instances, except for a few in which they were downplayed.

Extract 4 below is an example of how teachers reflect on content quality. In her evaluation of one of the students' definitions of "minaret," the teacher praises a student for having included different



Teacher D

The art answers [on 'colour and perspective'] had very little content, minimal.

Teacher F

And in those that had little content, the language had a lot of weight in the higher final 5. And in this one [about the Roman Empire] that has a lot of content, language has less weight, that's my feeling.

Extract 3.

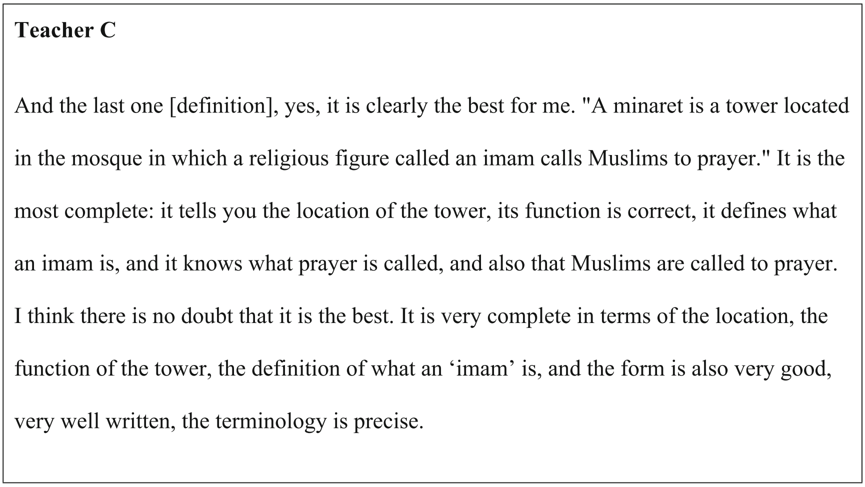
TABLE 6
Basis of Achievement Invoked for Content Knowledge

Content Performance	N = 112	%
Content quality	94	83.93
Content quantity	18	16.07
Content quality aspects		
Emphasized (QUALITY ER+)	90	80.36
Deemphasized (QUALITY ER−)	4	3.57
Content quantity aspects		
Emphasized (QUANTITY ER+)	17	15.18
Deemphasized (QUANTITY ER−)	1	0.89

specifying features (e.g., location, function, a related religious figure) that characterize the defined term, thus rendering it as "complete" in her view.

Extracts 5 and 6 are examples of emphasis on content quantity (Quantity ER+), which, as mentioned, came up in the teachers' reflections on fewer occasions. In Extract 5, a teacher relays how she generally stresses the length of the expected answers in exam situations by urging the students to "develop" their answers and provide more or "sufficient" information. In Extract 6, a teacher states that the amount of information should be a reason for favoring a student's answer (E) over another that had received a higher ranking (G) as the latter revolved solely around a single point, contrary to the former.

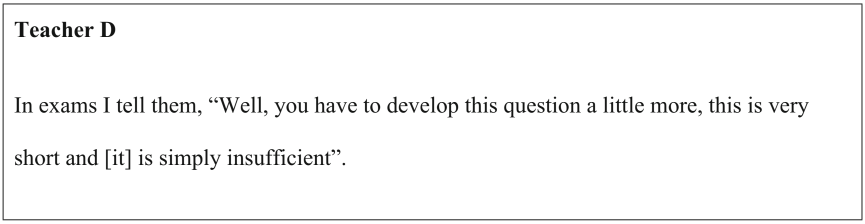
Extract 7 is an example of the discursive context in which teachers downplayed content, which is particularly uncommon for teachers to do. The exchanges refer to a student who was exploring the effects of



Teacher C

And the last one [definition], yes, it is clearly the best for me. "A minaret is a tower located in the mosque in which a religious figure called an imam calls Muslims to prayer." It is the most complete: it tells you the location of the tower, its function is correct, it defines what an imam is, and it knows what prayer is called, and also that Muslims are called to prayer. I think there is no doubt that it is the best. It is very complete in terms of the location, the function of the tower, the definition of what an 'imam' is, and the form is also very good, very well written, the terminology is precise.

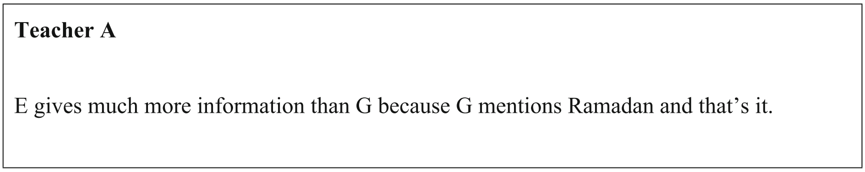
Extract 4.



Teacher D

In exams I tell them, "Well, you have to develop this question a little more, this is very short and [it] is simply insufficient".

Extract 5.

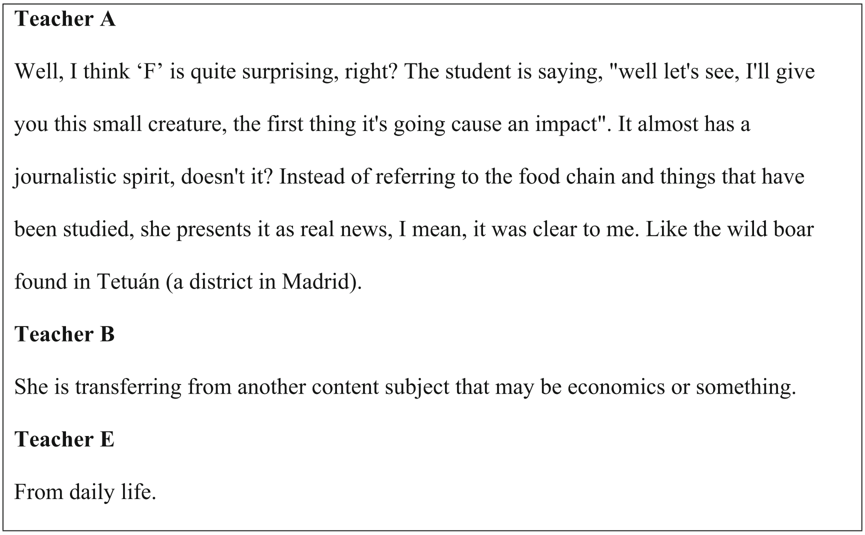


Teacher A

E gives much more information than G because G mentions Ramadan and that's it.

Extract 6.

introducing a new animal into a given ecosystem. Instead of drawing on the lexis and content from the textbook unit on ecosystems, the student wrote that introducing a platypus (the student's choice) would bring more tourists. The student's performance coincided with a news story about a wild boar found strolling in a Madrid neighborhood during the COVID-19 lockdown.



Teacher A

Well, I think ‘F’ is quite surprising, right? The student is saying, "well let's see, I'll give you this small creature, the first thing it's going cause an impact". It almost has a journalistic spirit, doesn't it? Instead of referring to the food chain and things that have been studied, she presents it as real news, I mean, it was clear to me. Like the wild boar found in Tetuán (a district in Madrid).

Teacher B

She is transferring from another content subject that may be economics or something.

Teacher E

From daily life.

Extract 7.

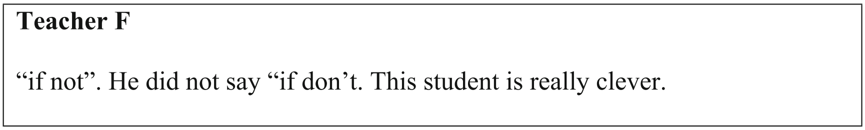
TABLE 7
Basis of Achievement Invoked for Language Knowledge

Language Performance	*N* = 95	%
Language form	64	67.37
Language function	31	32.63
Language form		
Emphasized (FORM ER+)	48	50.53
Deemphasized (FORM ER−)	16	16.84
Language function		
Emphasized (FUNCTION ER+)	27	28.42
Deemphasized (FUNCTION ER−)	04	04.21

Bases of Achievement for Language Knowledge

As shown in Table 7, roughly two-thirds of the coded instances involved the invocation of language form as a basis of achievement. Given that the seminar itself was designed to introduce the teachers to a functional model of language use (CDFs), it is unsurprising that language functions are less represented than language form, but it was unexpected to see a considerable number of instances (>15%) in which language form was downplayed (FORM ER−).

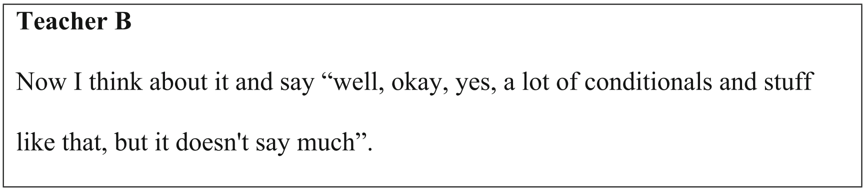
Extract 8 is an example of a language teacher emphasizing language form (FORM ER+) as a basis of achievement. The teacher praises a



Teacher F

“if not”. He did not say “if don’t. This student is really clever.

Extract 8.



Teacher B

Now I think about it and say “well, okay, yes, a lot of conditionals and stuff like that, but it doesn't say much”.

Extract 9.

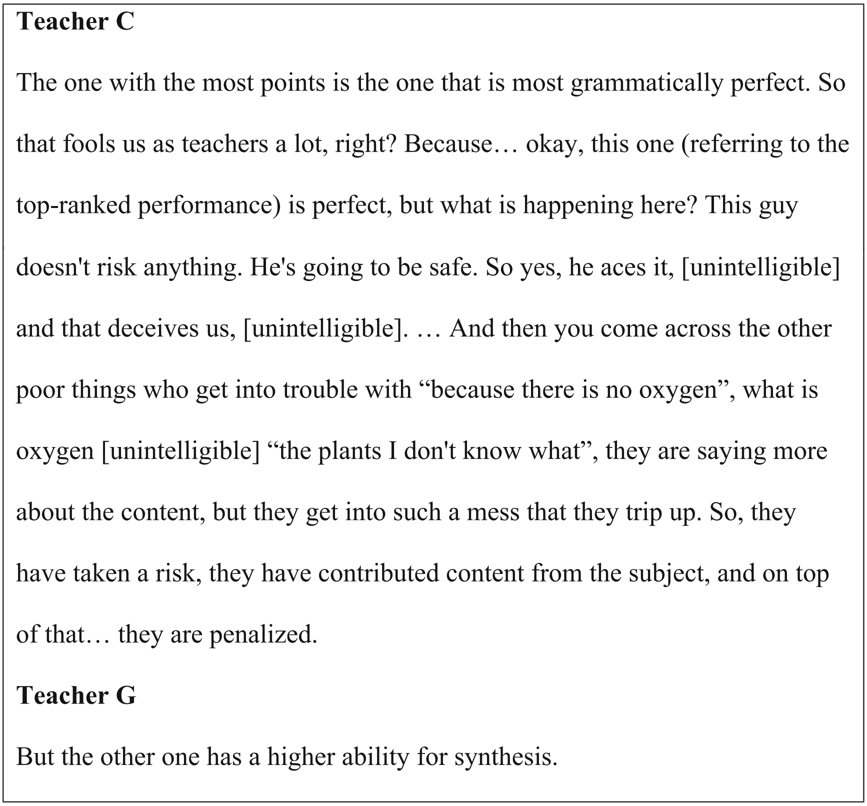
student who used the structure “because if not” to justify why it is important to protect the environment, which in her view was exceptionally clever. It is worth noting that the performance in question here was also considered of high content quality. In contrast, Extract 9 is an example of a history teacher who downplays the value of using linguistically complex structures (FORM ER−) where the overall content was weak. This reflects the primacy of content quality and meaning-making when evaluating students’ performance.

Extract 10 shows another context in which a teacher deemphasizes language form (FORM ER−). Here, two performances that communicate the same content information are contrasted. One is a shorter answer where synthesis of concepts was evident and which was ranked the highest, and the other is a longer answer where the student provided more content but had more redundancies and punctuation errors. According to teacher C, the second one received a seemingly unfair low ranking (rank = 3).

Although it was relatively rare for teachers to invoke language function as a basis of achievement (ER + LANG FUNCT), an example can be seen in Extract 11, where the teacher critiques the use of the non-formal definitional resource (“is like a”) and explicitly states that a more canonical academic structure is required.

Social Relations Invoked in Relation to Bases of Achievement

We coded 43 instances of invocations of students’ attributes (SR+) as bases of achievement in the teachers’ reflections (Table 8). There



Teacher C

The one with the most points is the one that is most grammatically perfect. So that fools us as teachers a lot, right? Because… okay, this one (referring to the top-ranked performance) is perfect, but what is happening here? This guy doesn't risk anything. He's going to be safe. So yes, he aces it, [unintelligible] and that deceives us, [unintelligible]. … And then you come across the other poor things who get into trouble with “because there is no oxygen”, what is oxygen [unintelligible] “the plants I don't know what”, they are saying more about the content, but they get into such a mess that they trip up. So, they have taken a risk, they have contributed content from the subject, and on top of that… they are penalized.

Teacher G

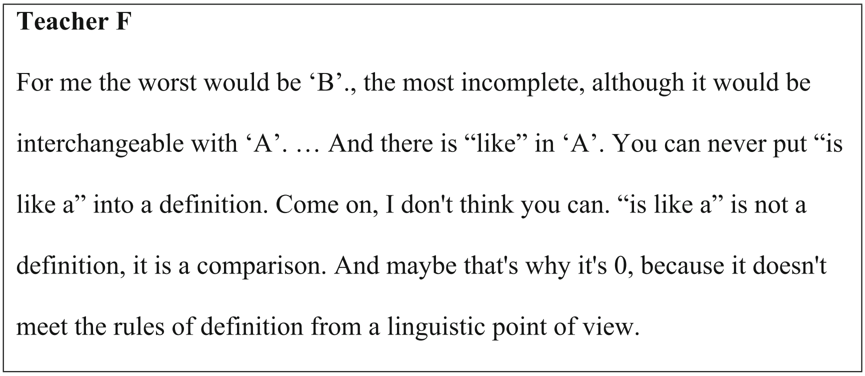
But the other one has a higher ability for synthesis.

Extract 10.

was an overall even balance between individual psychological factors (e.g., effort, willingness, critical thinking, criticality) and societal factors (e.g., culture, families, educational experiences).

Extract 12 is an example of a teacher invoking students’ ability to reason as a personal attribute (IND SR+) and this is seen as ultimately influencing how they use language to answer the prompt. Though the answers that were judged by the teachers were all from the same grade level, the teacher attributes the reasoning in the higher ranked answers to greater maturity of students, perhaps similar to performances expected of older children.

Extract 13 is an example of how teachers invoke societal and educational factors and practices (SOCIETAL SR +) that may lead them to different value linguistic performances depending on the language background of the student.



Teacher F

For me the worst would be 'B'., the most incomplete, although it would be interchangeable with 'A'. … And there is "like" in 'A'. You can never put "is like a" into a definition. Come on, I don't think you can. "is like a" is not a definition, it is a comparison. And maybe that's why it's 0, because it doesn't meet the rules of definition from a linguistic point of view.

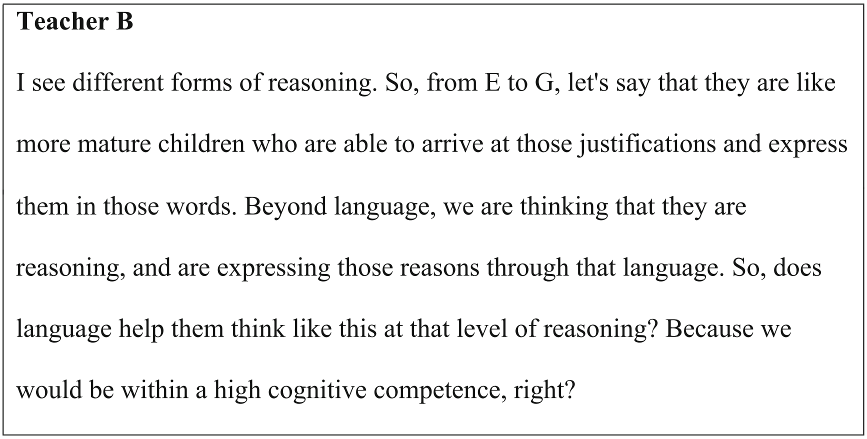
Extract 11.

TABLE 8
Social Relations Invoked in Relation to Achievement When Reflecting on Evaluation Criteria (SR)

Student attributes	$N = 43$	%
Individual factors	23	53.49
Societal factors	20	46.51
Individual attributes		
Emphasized (IND SR+)	23	53.49
Deemphasized (IND SR+)	0	0.00
Societal attributes		
Emphasized (SOCIETAL SR+)	19	44.19
Deemphasized (SOCIETAL SR−)	1	2.33

DISCUSSION AND CONCLUSIONS

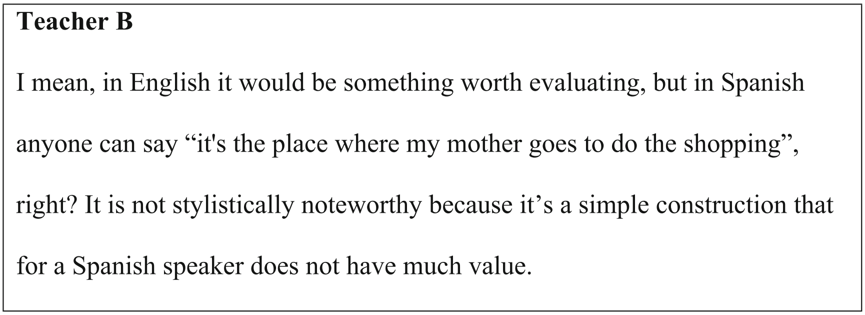
Overall, the results of the study suggest that the teachers were oriented mostly to epistemic relations (ER) in articulating the bases of achievement after judging the samples of students' work, with social relations (SR) accounting for 17% of the total. This would suggest that, at least as expressed in the teachers' reflections, a knowledge code is operating in this bilingual education practice. The teachers seem to emphasize the objects and aspects of reality the practice is oriented to, both in terms of disciplinary content, and language. They less often explicitly consider social relations, that is, who the learners are in relation to the practice. Within epistemic relations, the fact that there is an overall balance between the two types of knowledge, content and language, suggests that the ground is fertile for the establishment and maintenance of CLIL practices, which entail a dual focus on



Teacher B

I see different forms of reasoning. So, from E to G, let's say that they are like more mature children who are able to arrive at those justifications and express them in those words. Beyond language, we are thinking that they are reasoning, and are expressing those reasons through that language. So, does language help them think like this at that level of reasoning? Because we would be within a high cognitive competence, right?

Extract 12.



Teacher B

I mean, in English it would be something worth evaluating, but in Spanish anyone can say “it's the place where my mother goes to do the shopping”, right? It is not stylistically noteworthy because it’s a simple construction that for a Spanish speaker does not have much value.

Extract 13.

both types of knowledge in instruction and assessment. The data suggest that content teachers are open to including language criteria in assessing students’ work, and language teachers see the importance of the quality of content knowledge displayed by the students. Indeed, there is evidence that some of the teachers have a sophisticated professional understanding of the delicate balancing act involved in judging the quality of student work when it is written in the L2. They are aware that they can be dazzled by impressive language skills, which may hide gaps in content understanding, or that some students’ work may be unjustly penalized for surface language errors when it shows a good understanding of content.

However, the fact that social relations were explicitly invoked 43 times (17%) does not wholly account for the complex relationships between social and epistemic relations in these teachers’ assessment

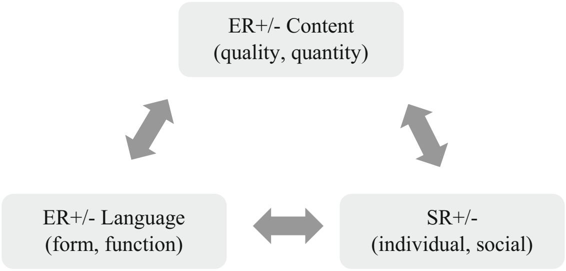


FIGURE 2. Relationships between social and epistemic relations as bases of achievement.

practices. The prevalence of invocation of a linguistic form as a basis of achievement over language function (48/27 occasions coded) also has implications for the importance of social relations. In order to tease out the interconnections between epistemic and social relations as evinced in the criteria invoked by the teachers in the study, it may be useful to present them in graphic form, as a basis for discussion (Figure 2).

When language form is emphasized over functional aspects of language directly related to the content knowledge, the bases of achievement do not directly reflect what has been taught to the students during instruction, and thus there is an increased risk of construct-irrelevant variance contaminating the teachers' judgment of the extent and quality of content learning (Avenia-Tapper & Llosa, 2015), as seen in the double arrow on the left-hand side of the figure. The nonrelevant aspects of performance, which reflect general language proficiency factors such as fluency and accuracy can be the result of experiences gained outside the current learning experience (societal factors) or may reflect individual language aptitude or motivation (personal attributes), as seen in the bottom double-headed arrow. In this sense, highlighting aspects of general language proficiency invokes a knower code, in which legitimacy may reside in being the type of person who has relevant attributes and/or experience.

A knower code can also manifest in relation to content knowledge (right-hand double-headed arrow), and in this case, a strong emphasis on social relations in relation to the expression of content knowledge is more likely to be attributed to individual psychological factors (IND SR+). This occurs when the bases for achievement invoked are students' creativity, originality, inventiveness, etc. Such a knower code orientation is not necessarily problematic from a CLIL perspective as it may relate to pedagogical practices within the content subject being taught, or institutional and cultural educational preferences (e.g., an emphasis on constructivism or "student-centered" pedagogy). Whatever the debates about the merits of such approaches in the wider educational sphere, provided there is alignment between instruction and assessment practices (Lo, Lui, Wong, 2019), the scaffolding of the functional language required to complete learning tasks (Lo & Fung, 2020), and assessment tasks are balanced between cognitive and linguistic demand (Lo & Lin, 2014; Shaw & Imam, 2013), such an orientation is not problematic from the point of view of content and language integration. Indeed, it is possible to envisage a CLIL practice that would be a knower code in both content and language (creativity, use of own experience in the content area, and language skills "brought along" from previous experiences and/or individual aptitude). Where both epistemic and social relations are emphasized, it is possible to envisage an "élite" CLIL code, which would place a high premium on content knowledge but would require the successful student to be the "right kind of person," perhaps with individual gifts and talent in the content area, and polished language skills gained through privileged access to the L2. Thus, though élite codes in the LCT dimension of Specialization do not refer to social exclusivity (Maton, 2014, p. 31), there is a sense in which, in the context of bilingual education, when the connection between the epistemic relations in terms of content and language and social relations in terms of previous experience is taken into account, élite takes on both meanings. Such an orientation in any CLIL context is likely to involve assessment practices that risk rewarding accomplished linguistic performance and downgrading the work of students who may have adequate content knowledge but lack polished language skills in ways that are not functional for the expression of content knowledge.

In order to overcome the "blind spots" in CLIL assessment, it may be beneficial to consider the possibility of CLIL orienting to a knowledge code, for both content and (functional aspects of) language. This would be particularly appropriate for educational cultures where a knowledge code orientation exists in the content areas, as is suggested by the data in the current study. This would reduce the risk of "code clash"(Maton, 2014, p. 73), in which the specialization code of the content area did not match that of the approach to language. In this approach, CLIL instruction and assessment would be based on the explicit teaching of the language which is functional for the communicative tasks related to content learning. It would reflect the "higher disciplinary orientation to language/more visible language pedagogy" orientation described by Leung and Morton (2016) and be consistent with calls in the literature for a functional approach to language for instruction and assessment (Avenia-Tapper & Llosa, 2015; Chadwick, 2012; Otto, 2018). In this way, language would no longer be an "invisible" component in CLIL assessment (Hönig, 2010), and threats to validity, fairness, and equity, in which students are assessed on what they have not been explicitly taught, would be reduced.

Implementing a "knowledge code" orientation to CLIL instruction and assessment is likely to be a challenging task for teacher preparation and professional development. There is evidence in the literature that content teachers in bilingual education contexts find it difficult to see themselves as responsible for their students' language development (Cammarata & Tedick, 2012; Tan, 2011), and even language teachers find it challenging to identify language objectives relevant to content learning (Baecher, Farnsworth, & Ediger, 2014). The findings of the current study suggest that the teachers' perception of what "language" is in this bilingual education context is mostly a matter of language structures and forms, which can be an object of assessment in isolation from the content knowledge which is the focus of instruction. There is less evidence of a focus on language which is functional for specific learning outcomes or instructional tasks.

The study's results lend support to the argument that professional development within CLIL contexts should seek to increase teachers' awareness of functional aspects of language use in relation to the learning of academic subjects. Doing so would enable them to apply more integrative assessment criteria, which they could also share with students and incorporate into their teaching. It would also contribute to ensuring equity, as gaining control of academic language functions is a challenge for all learners, irrespective of language background. Shifting the emphasis away from isolated language forms would avoid giving an unjust advantage to those students who have had the opportunity through familial and/or economic circumstances to have had a greater acquaintance with the L2 medium of instruction.

In sum, such work would have the intended outcome of developing in teachers a kind of assessment literacy for CLIL, which would enable them to explicitly identify language objectives functional for content learning tasks and outcomes, provide scaffolding for these objectives in instruction, and align instructional and assessment tasks. This requires teachers to have access to robust, research-tested models for content and language integration in preservice education and professional development, and thus have a shared language for bases of achievement. Such models, such as Lo and Lin's (2014) matrix, Coyle and Meyer's (2021) pluriliteracies, DeBoer and Leontjev's (2020) classroom-based assessment framework, and Dalton-Puffer's (2013) cognitive discourse functions are available. What remains is for them to be "normalized" in teachers' practices (Coyle, 2018) through the provision of pre- and in-service teacher education and the availability of materials that reflect a truly integrated pedagogy. This is increasingly urgent as CLIL continues to spread throughout the world. Failure to tackle the blind spot of assessment in CLIL may put at risk social equity by potentially harming the educational prospects of more and more learners who are expected to study academic content in an additional language, which is usually English.

FUNDING

This research was funded by the Spanish Ministry of Science, Innovation and Universities (LongAd-CLIL project-RTI2018-094961-B-I00).

THE AUTHORS

Tom Morton is a Beatriz Galindo Distinguished Research Fellow in the Department of English Studies, Universidad Autónoma de Madrid, Spain, where he is a member of the UAM-CLIL Research Group. His research interests include Content and Language Integrated Learning (CLIL), English-medium instruction (EMI), classroom discourse, and language teacher knowledge and identity.

Nashwa Nashaat-Sobhy is an Associate Professor in Applied Linguistics at Universitat Politècnica de València. She is a research member of "GALE" and "UAM-CLIL Group." Her research centers mainly on teaching and learning in English-Medium Instruction contexts, on which she has published in John Benjamins, Taylor & Francis, and Routledge, among others.

REFERENCES

Avenia-Tapper, B., & Llosa, L. (2015). Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English language learners' science knowledge. *Educational Assessment, 20*(2), 95–111.

Baecher, L., Farnsworth, T., & Ediger, A. (2014). The challenges of planning language objectives in content-based ESL instruction. *Language Teaching Research, 18*(1), 118–136.

Bauer-Marschallinger, S. (2022). *CLIL with a capital I: Using cognitive discourse functions to integrate content and language learning in CLIL history education.* (Publication No. UA792343) [Doctoral dissertation, Vienna University].

Cammarata, L., & Tedick, D. J. (2012). Balancing content and language in instruction: The experience of immersion teachers. *The Modern Language Journal, 96* (2), 251–269.

Chadwick, T. (2012). *Language awareness in teaching: A toolkit for content and language teachers.* Cambridge, England: Cambridge University Press.

Codó, E. (2022). *Global CLIL: Critical, ethnographic and language policy perspectives (Critical studies in multilingualism).* London, England: Routledge.

Coyle, D. (2018). The place of CLIL in (bilingual) education. *Theory Into Practice, 57*(3), 166–176.

Coyle, D., & Meyer, O. (2021). *Beyond CLIL: Pluriliteracies Teaching for Deeper Learning*. Cambridge, England: Cambridge University Press.

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., Hüttner, J., & Llinares, A. (2022). CLIL in the 21st Century: Retrospective and prospective challenges and opportunities. *Journal of Immersion and Content-Based Language Education, 10*(2), 182–206.

de Boer, M., & Leontjev, D. (Eds.). (2020). *Assessment and learning in content and language integrated learning (CLIL) classrooms: Approaches and conceptualisations.* Cham, Switzerland: Springer.

He, P., & Lin, A. M. (2019). Co-developing science literacy and foreign language literacy through "Concept+ Language Mapping". *Journal of Immersion and Content-Based Language Education, 7*(2), 261–288.

Hidalgo-McCabe, E. A., & Fernández-González, N. (2019). Framing "choice" in language education. In L. M. Rojo & A. Del Percio (Eds.), *Language and neoliberal governmentality* (pp. 68–90). London, England: Routledge.

Hönig, I. (2010). *Assessment in CLIL: Theoretical and Empirical Research.* Saarbrücken: VDM Verlag Dr. Müller.

Leung, C., & Morton, T. (2016). Conclusion: Language competence, learning and pedagogy in CLIL - deepening and broadening integration. In T. Nikula, E. Dafouz, P. Moore & U. Smit (Eds.), *Conceptualising integration in CLIL and multilingual education* (pp. 235–248). Bristol, England: Multilingual Matters.

Llinares, A., & Evnitskaya, N. (2021). Classroom interaction in CLIL programs: offering opportunities or fostering inequalities? *TESOL Quarterly, 55*(2), 366–397.

Lo, Y. Y., & Fung, D. (2020). Assessments in CLIL: The interplay between cognitive and linguistic demands and their progression in secondary education. *International Journal of Bilingual Education and Bilingualism., 23*(10), 1192–1210.

Lo, Y. Y., Fung, D., & Qiu, X. (2021). Assessing content knowledge through L2: mediating role of language of testing on students' performance. *Journal of Multilingual and Multicultural Development*, 1–16.

Lo, Y. Y., & Lin, A. (2014). Designing assessment tasks with language awareness: Balancing cognitive and linguistic demands. *Assessment and Learning, 3*(3), 97–119.

Lo, Y. Y., Lui, W. M., & Wong, M. (2019). Scaffolding for cognitive and linguistic challenges in CLIL science assessments. *Journal of Immersion and Content-Based Language Education, 7*(2), 289–314.

Martín-Rojo, L. (2013). Capitalising students through linguistic practices: A comparative analysis of new educational programmes in a global era. In A. Duchêne, M. Moyer & C. Roberts (Eds.), *Assessing content knowledge through L2: mediating role of language of testing on students' performance* (pp. 118–146). Bristol, England: Multilingual Matters.

Massler, U., Stotz, D., & Queisser, C. (2014). Assessment instruments for primary CLIL: The conceptualisation and evaluation of test tasks. *The Language Learning Journal, 42*(2), 137–150.

Maton, K. (Ed.). (2014). *Knowledge and Knowers. Towards a realist sociology of education.* London, England: Routledge.

Maton, K., & Chen, R. T.-H. (2016). LCT in qualitative research: Creating a translation device for studying constructivist pedagogy. In K. Maton, S. Hood, & S. Shay (Eds.), *Knowledge-building: Educational studies in Legitimation Code Theory* (pp. 27–48). London, England: Routledge.

Mohan, B., Leung, C., & Slater, T. (2010). Assessing language and content: A functional perspective. In A. Paran & L. Sercu (Eds.), *Testing the untestable in language education* (pp. 217–240). Bristol, England: Multilingual Matters.

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.

Morton, T., & Llinares, A. (2017). Content and language integrated learning (CLIL): Type of programme or pedagogical model? In A. Llinares, & T. Morton (Eds.), *Applied linguistics perspectives on CLIL* (pp. 1–16). Amsterdam, the Netherlands: John Benjamins.

O'Donnell, M. (2021). UAM CorpusTool version 3.3. Retrieved from http://www.corpustool.com/index.html

Otto, A. (2018). Assessing language in content and language Integrated learning: A review of the literature towards a functional model. *Latin American Journal of Content & Language Integrated Learning, 11*(2), 308–325.

Otto, A., & Estrada, J. L. (2019). Towards an understanding of CLIL assessment practices in a European context: Main assessment tools and the role of language in content subjects. *CLIL Journal of Innovation and Research in Plurilingual and Pluricultural Education, 2*(1), 31–42. https://doi.org/10.5565/rev/clil.11

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education, 84*, 128–138.

Pérez Cañado, M. L. (2020). CLIL and elitism: myth or reality? *The Language Learning Journal, 48*(1), 4–17. https://doi.org/10.1080/09571736.2019.1645872

Shaw, S., & Imam, H. (2013). Assessment of international students through the medium of english: ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly, 10*(4), 452–475.

Tan, M. (2011). Mathematics and science teachers' beliefs and practices regarding the teaching of language in content learning. *Language Teaching Research, 15* (3), 325–342.

Wheadon, C., Pinot de Moira, A., & Christodoulou, D. (2020). The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment. https://doi.org/10.31235/osf.io/vzus4