openjournals.nl

Dutch Journal of Applied Linguistics
https://www.dujal.nl | E-ISSN 2211-7253

Integration, demands and accommodations: An exploration of summative assessment practices in Dutch bilingual secondary education

Tessa Mearns[1], Catherine van Beuningen[2,3], Niels Nederlof[4] and Nivja H. de Jong[1,5]
[1] *ICLON Leiden University Graduate School of Teaching* | [2] *Amsterdam University of Applied Sciences* | [3] *University of Amsterdam* | [4] *Dialogic* | [5] *Leiden University Centre for Linguistics*

Abstract In content and language integrated learning (CLIL), the goal is to support learning of both subject content and (subject-related) language, through an integrated approach. Based on the principle of constructive alignment (Biggs, 1996), it is logical to assume that such integration carries through to assessment. This study explores the extent to which this is the case in the context of bilingual secondary education in the Netherlands. Drawing on quantitative, self-report data from a teacher questionnaire, and qualitative analysis of assessment materials, we explore the relative roles of language and content, the use of accommodations and translanguaging as support, and the cognitive and linguistic demands posed by assessments. In line with previous research, it appears that assessment practices do not always align with the integrated goals of CLIL. Implications for practice are discussed, as is the need for more attention for this aspect of CLIL in research and teacher education.

Keywords content and language integrated learning (CLIL), assessment, summative, teaching practice, bilingual education, accommodations, translanguaging, cognitive demands, linguistic demands, integration

Article history
Received: July 6, 2023
Accepted: November 6, 2024
Online: May 21, 2025

Corresponding author
Tessa Mearns,
t.l.mearns@iclon.leidenuniv.nl

Author contributions
Funding information
Statement of interest
Statement of technology use
Supporting information
See p. 25–26

1 Introduction

Teachers in content and language integrated learning (CLIL) contexts are faced with a dual teaching task (Coyle et al., 2010; Dalton-Puffer et al., 2014). They have to guide learners in their content learning, and in their development of second language (L2) proficiency, in particular in subject-specific discourses (Coyle & Meyer, 2021). Following the principle of constructive alignment (Biggs, 1996), it seems logical to assume that this dual, integrated teaching and learning goal will carry through to the way in which learning is assessed (Sato, 2023). In other words, if language and content are integrated in CLIL instruction, one could expect them to also be integrated in CLIL assessment.

As the limited research on summative assessment in CLIL (e.g. Otto, 2017) has shown, however, this assumption may not be so self-evident. Indeed, there appears to be a lack of consensus regarding the extent to which language and content in CLIL should be

How to cite Mearns, T., Van Beuningen, C., Nederlof, N. and De Jong, N.H. (2025). 'Integration, demands and accommodations: An exploration of summative assessment practices in Dutch bilingual secondary education', *Dutch Journal of Applied Linguistics* 14: DOI 10.51751/dujal15756

assessed separately or in combination. While practical publications aimed at CLIL teachers generally agree that the question of content and language integration adds extra complexity to assessment, recommendations as to how to respond to this complexity are not always clear-cut or consistent. As Lo and Fung (2020) rightfully argue, without a clear framework guiding CLIL teachers in the design of assessments, we run the risk of assessing learners in unfair and invalid ways.

In the Netherlands, which has one of the most well-established and clearly regulated paradigms of bilingual education in Europe (Mearns & de Graaff, 2018), the picture is equally murky. While Dutch secondary education in general is heavily focused on summative assessment (Sluijsmans et al., 2013), the national Quality Standard for Bilingual Education (Nuffic, 2019) makes no mention of how this aspect of CLIL should or could be addressed, and there has been no research published on assessment practices in this setting (Mearns et al., 2023). Therefore, we have a limited picture of how CLIL teachers in the Netherlands approach assessment of language and content. The current, small-scale study therefore aims to map how these teachers shape their assessment practices, in absence of clear guidelines, and as such to spark discussion regarding the ways in which learning is assessed in this and other multilingual education contexts. Moreover, as language plays a crucial role in any kind of learning, irrespective of the content being learned or the language in which the learning takes place in, we hope to open a similar discussion regarding the role of language in assessment in mainstream education.

1.1 Separate vs. integrated assessment

As CLIL encompasses the teaching of both content and language, it follows that valid CLIL assessment will also be dual-focused (Llinares et al., 2012; Mehisto & Ting, 2017; Otto, 2017; Lo & Fung, 2020). What is less clear, however, is whether this means that the content and language taught and learned through CLIL should – and can – be assessed together or separately (Otto, 2019).

Some practical publications (e.g. Bentley, 2010; Coyle et al., 2010; Dale & Tanner, 2012) aim to help CLIL teachers identify the focus of assessment and to decide whether the emphasis should be on content, language or other related skills. They suggest that distinctions might also be made between formative and summative assessment, whereby content assessment alone is linked to grading, but feedback is provided on language. Other authors have questioned whether the separation of language and content in assessment is desirable considering the integrated goals of CLIL (e.g., Llinares et al., 2012; Mehisto & Ting, 2017; Morton, 2020). They suggest that language and content should be assessed together, as language is an integral aspect of subject content.

Calls for integrated assessment reflect recent thinking on CLIL from a disciplinary literacies perspective, which highlights that the content, language and cognitive processes involved in a subject are inextricably linked (e.g., Dalton-Puffer et al., 2018; Coyle & Meyer, 2021), or, as Morton (2020) formulates it: “Learning a school subject [...] means being able to comprehend and produce the types of texts or genres (both oral and written) through which knowledge in the subject is communicated" (p. 9). In an integrated model of CLIL assessment (Llinares et al., 2012), this inextricable relationship between content and language is acknowledged at all stages of the learning process. The tasks and subject-specific language used in assessment will reflect those encountered in the classroom, so that learners know what is expected of them and how they will be evaluated.

Sato (2023) addresses the variety of perspectives on assessment in CLIL by focusing on the purposes of assessment. He discerns three types of CLIL assessment: 'separate', 'weakly integrated' and 'strongly integrated'. Sato emphasizes that, while separation of language and content (e.g. through standardized language tests and content-focused testing in the first language) can be useful for research purposes, it does not align with the integrated goals of CLIL and is therefore a less logical choice for assessment in teaching practice. Weakly integrated assessments mitigate potential interference from lower L2 proficiency by managing linguistic demands, for example through contingent scaffolding during oral interaction or reducing the amount of active language production required in written tasks. This, Sato suggests, can be useful for diagnostic or formative assessment of content learning. For summative assessment that is aligned with CLIL's integrated goals, however, Sato recommends a strongly integrated approach in which language is regarded as an aspect of the subject content. He argues that, if teaching introduces and supports the learning of subject-specific language as integral to subject learning, it follows that assessment should also be integrated.

1.2 Balancing cognitive and linguistic demands

While Lo and Fung (2020) agree that integrated assessment approaches are the best fit with the dual goal of CLIL, they also point to the fact that "CLIL assessments should avoid underestimating students' content knowledge due to their inability to express their understanding with appropriate language" (p. 1194). Whereas Sato (2023) proposed to mitigate this interference between language proficiency and content-related performance in diagnostic and formative assessment only (i.e. 'weak integration'), others suggest that support or 'accommodations' can also help increase the validity of summative assessment (Coyle et al., 2010; Lo & Lin, 2014; Otto, 2017) by "[helping] students access the content in English and better demonstrate what they know" (Butler & Stevens, 1997, p. 5). As defined by researchers in non-CLIL multilingual education settings, accommodations can be embedded in the assessment format (e.g. visual aids, text modifications, glossaries, translations) or made available as part of the assessment procedure (e.g. providing explanations on request; access to support materials such as dictionaries or online materials) (De Backer et al., 2019a; Yang, 2020). Accommodations have been identified as both helping learners and strengthening the validity of assessments (De Backer, 2020; Lo & Lin, 2014), although which types of accommodations are most effective appears to vary between learners and situations. De Backer et al. (2019a) therefore recommend the use of a range of accommodation strategies. Examples of accommodations that have been identified as both helpful to learners and beneficial in strengthening the validity of assessments are adaptations of texts to make them more accessible; providing aural input alongside written texts; and providing access to support materials such as glossaries or dictionaries (De Backer et al., 2019a; Lo & Lin, 2014).

A specific accommodation that can be used to balance the linguistic and cognitive assessment demands, is 'translanguaging': providing room for learners to use multiple languages in assessment and/or response formats (De Backer et al., 2019a; De Backer et al., 2019b). Translanguaging can support the learning of both content and language (Cenoz, 2017) and has been found to be an effective tool in minimizing interference of language in the assessment of subject content (Serra, 2007). In this sense, L1 support could be a helpful tool in what Sato (2023) terms 'weakly integrated' CLIL assessment where the main focus is on content outcomes. It is important to note, however, that learners do not always perceive choice of language in assessments as helpful, as identified in both CLIL and other L2 content classrooms (De Backer, 2020; Hönig, 2010). As De Backer et al. (2019a) observed, individual factors such as L2 proficiency and personal preference can influence the extent to which learners perceive translanguaging in assessments as useful.

1.3 Previous research on CLIL assessment practices

Research on assessment practices in CLIL contexts is scarce. From the few studies that are available, it seems that CLIL teachers have a preference for separating language and content in assessments, but that they do not always achieve this in practice. Hönig (2010), for example, observed a contrast between teachers' conviction that they focused only on content in CLIL assessment, and their actual assessments, which appeared to be influenced by linguistic aspects such as fluency and length of responses. Similarly, Otto and Estrada (2019) observed that content teachers found it difficult to ignore language in assessment, but at the same time did not feel equipped to assess it. These studies highlight how language can be an "invisible" criterion in CLIL assessment (Llinares et al., 2012, p. 284). Furthermore, the preference among CLIL teachers to take the same approaches to assessment with CLIL groups as they do with non-CLIL groups (Otto, 2017; Reierstam & Sylvén, 2019), namely with little attention for language, also seems to suggest that CLIL teachers may be insufficiently aware of the role language plays in assessment. According to Lo et al. (2019), an invisible role for language in CLIL assessment is not only a threat to fairness and validity, but can also affect the extent to which these aspects are integrated in CLIL teaching. They noted that a lack of explicit attention for language in assessment can have a washback effect, leading teachers to neglect the linguistic aspects of their subject in classroom teaching.

Multiple studies from CLIL contexts have shown that lack of L2 proficiency can hinder CLIL learners in the expression and demonstration of their content knowledge and skills, especially in the earlier stages of L2 development. Oattes et al. (2020) observed that learners in their first year of CLIL performed better in a multiple-choice history assessment conducted in Dutch (the main school language) than in a comparable assessment conducted in English (the CLIL language). This effect was not observed among CLIL learners in their third year, suggesting that learners with higher L2 proficiency coped better with the assessment's demands. Likewise, Lo and Fung (2020) found that students' performance in CLIL assessments declined with increasing linguistic demands. The latter study furthermore showed that, in CLIL assessments for science and biology in upper secondary education, both the cognitive and linguistic demands were generally high. In light of their findings pertaining to students' performance, the authors warn that this situation is problematic as the high linguistic demands may prevent students with lower L2 proficiency from demonstrating their understanding of subject content. This also reflects evidence from non-CLIL contexts where limited mastery of the instructional language was found to negatively influence students' performance on content exams (Trenkic & Warmington, 2019).

Little CLIL research has explored the potential of accommodations to mitigate interference between language and content in assessment. Regarding the presence of accommodations in CLIL assessment, while around two-thirds of the assessment materials analysed by Otto (2017) contained visual or pictorial support to limit language demands, only a quarter used graphic organizers to support language production, and none made use of glossaries or other types of linguistic support. To the best of our knowledge, there are no studies that explore if and how translanguaging can support CLIL assessment. This might be related to the fact that in some CLIL contexts, including the Netherlands, there is a strong conviction among schools and teachers that communication in CLIL should take place exclusively in the CLIL target language (Oattes et al., 2018), based on widespread misconceptions about multilingual language development (van Beuningen & Polišenská, 2019). That said, Hönig (2010) observed that, when students were offered a choice of language in which to complete an assessment, they all opted for the language in which they had been taught (English) rather than the main school language (German). This could corroborate Coyle et al.'s (2010) argument that it is inconsistent to assess learners in a language other than the one in which they have been taught. Considering the arguments presented above from other, non-CLIL, multilingual contexts, however (e.g. De Backer et al., 2019a), it seems that offering L1 translations or other materials as optional support during content-focused assessments is potentially helpful for some learners.

1.4 Current study

As explored above, based on the principle of constructive alignment (Biggs, 1996) and on the theoretical recommendations for CLIL practice, it could be assumed that (a) the integrated goals of language and content learning would translate into integrated approaches to assessment in CLIL contexts; and (b) scaffolded approaches as used in

CLIL teaching would feed into the use of accommodations and translanguaging in assessment, to balance the cognitive and linguistic demands of assessment tasks. From the scarce research carried out in CLIL settings so far, however, it appears that practices in both of these regards are varied and that teachers are not always aware of the interaction between content and language in the way they assess learners' performance. With this in mind, the current small-scale study explored ways in which summative assessment is approached in the first three years of Dutch-English bilingual secondary education (BSE) in the Netherlands, with the aim of increasing our understanding of CLIL assessment in this specific context.

Whereas CLIL is often associated with teaching and learning in non-language subjects, such as biology, history or art, many teachers of English in bilingual education (TEBs) in the Netherlands identify both as language acquisition specialists, and as content specialists in literature and language arts (Dale et al., 2018). This could lead to TEBs and teachers of other subjects (STs) having different and possibly complementary perspectives on assessment, although we consider them all CLIL teachers. Therefore, although it was not the main goal of the current study to compare the assessment practices of STs and TEBs, we analysed these groups' practices separately.

The research question guiding this study is: How do STs and TEBs in bilingual secondary education in the Netherlands approach summative assessment of language and content?

The answer to this question was sought in relation to the themes discussed in the preceding section, namely the extent of content and language integration; the interplay between cognitive and linguistic demands; and the use of strategies to balance these demands (i.e. accommodations and translanguaging).

2 Method

To answer the research question, summative assessment practices in Dutch-English BSE were explored in both their perceived and operational forms (van den Akker, 2003), combining a broader examination of teachers' self-reported practices with in-depth analysis of assessment materials from actual teaching practice. This allowed us to explore and understand assessment practices from teachers' perspectives while also gaining a more objective impression of what those practices look like.

2.1 Research context

The research took place in the context of lower secondary (learners aged 12-15) BSE in the Netherlands. BSE (known locally as *tweetalig onderwijs* or *TTO*) has existed since 1989 and has followed a national quality standard and accreditation scheme since 2003. At the time of writing, 134 schools offer BSE, amounting to about 21% of the country's

648 secondary schools (Mearns et al., 2023). BSE schools generally follow the same curriculum as mainstream schools, but offer a significant proportion of that curriculum in English. For pre-vocational (or *vmbo*) learners, the minimum proportion of English-medium teaching is 30 % in years 1-2, and for learners in the higher general (or *havo*) and pre-university (or *vwo*) tracks it is 50 % in years 1-3. By the end of their third year in BSE, learners are expected to reach A2 (pre-vocational/*vmbo*), B1 (higher general/*havo*) or B2 (pre-university/*vwo*) level English (according to the Common European Framework of Reference for Languages, CEFR). English-medium provision is spread across different subjects, so that learners develop L2 proficiency in a range of disciplines. In the senior years, most subjects revert to teaching in Dutch, in preparation for the centralized final examinations. The English-as-discipline curriculum in the senior years of BSE is enhanced with a content-oriented international examination program, in most cases the International Baccalaureate (IB).

2.2 Participants, data collection and data-analysis

The study combined two stages of data collection: a survey aimed at painting a broad picture of teachers' (N = 42) self-reported summative assessment practices, and an in-depth analysis of example assessments (N = 17) submitted by thirteen teachers. The participants, instruments and analyses pertaining to each stage are described below.

2.2.1 *Stage 1: Self-reported assessment practices*

2.2.1.1 Participants

Participants were recruited using a convenience sampling method, via the Network of Dutch Bilingual Schools and the researchers' own networks. In total, 45 teachers teaching in the junior years of BSE from 17 different schools filled in the questionnaire completely (an additional five teachers started but did not complete the questionnaire). We excluded three teachers: one German language teacher, one teacher with no CLIL experience, and one teacher of physical education. Of the 42 participants, ten were TEBs. The remaining 32 STs taught a variety of subjects: history (6); mathematics (5); geography (6); art/music/drama (6); biology (4); physics/chemistry (3); economics (1); religious studies (1). The teachers had between 1 and 37 years' teaching experience and between 1 and 30 years' experience of teaching in BSE.

2.2.1.2 Questionnaire

An online questionnaire was designed to gather information on (1) teachers' summative assessment practices in general, (2) the relative role of language and content in assessments, and (3) how teachers used accommodations and translanguaging in assessment. The questionnaire was published in Formdesk and consisted of two sections. The first section requested background information regarding the teachers and their school context. The second section contained open and multiple-choice questions regarding the three themes specified above, including typical assessment types, integration of content and language in assessment, relative weighting of content and language in grading, language(s) used in assessment, and the availability of accommodations during assessment. This section also contained a matrix in which participants could indicate which aspects of language were addressed in grading and in the feedback they provided. The latter was included in the knowledge that summative assessments can also serve a formative function, so it may be the case that teachers give feedback on language aspects even when they do not grade them. For several questions, a distinction was drawn between summative 'tests' (i.e. "an assessment that is completed in class time and measures what a student can do at that moment") and summative 'assignments' (i.e. "products that a student or group of students might spend longer on, perhaps over a number of lessons or including time outside of class, such as an essay, short story, project, lab report, presentation, artwork or video"), as assessment practices might be different for each method, for example due to time constraints or the availability of help and resources. Therefore, handling all assessment types together might create a less clearly defined picture of current assessment practices than if tests and assignments were approached separately. The questionnaire can be found in Appendix S1.

2.2.1.3 Procedure

Participants received a link to the questionnaire by email or via social media. Before entering the questionnaire, participants were made aware of the fact that participation was voluntary and that their answers would be processed anonymously, and they were asked to give their consent. The final question asked if participants would be willing to share their materials (for the next stage of the data collection) and if so, to share an email address. Otherwise, participants were anonymous. Data were gathered in October-November of 2021. The time needed to fill in the questionnaire was around 15 minutes.

2.2.1.4 Analyses

Quantitative data gathered in the multiple-choice items were analyzed descriptively. Additional explanatory comments and additional responses to multiple-choice questions were categorized and are reported quantitatively. Responses from TEBs and STs are reported separately.

2.2.2 *Stage 2: Actual assessment practices*

2.2.2.1 Participants

In this stage too, teachers from years 1-3 of BSE were approached via the Network of Dutch Bilingual Schools and the researchers' professional networks, and through the final question of the Stage 1 questionnaire. This round of recruitment was separate to the invitation to complete the questionnaire. 34 teachers, including 12 questionnaire respondents, registered their interest in participating and were sent instructions as to how to do so. Thirteen teachers agreed to participate: two TEBs and eleven STs.

The spread of subjects can be found in Table 1. Participants had between 2 and 26 years' teaching experience, of which between 2 and 25 years in BSE. 6 of the teachers (with between 2 and 25 years' experience) had taught in BSE from the start of their career. Two teachers self-reported to be 'native' speakers of English; two had university degrees in English; seven reported having obtained a Cambridge Proficiency in English (CPE) certificate (CEFR level C2); two reported having no formal language certification, but estimated their own level of English at C2 or "excellent". All participants reported having received some CLIL training, either as professional development in school or externally, or as part of their initial teacher education.

2.2.2.2 Assessment bundles

Participating teachers were asked to share one or more 'assessment bundle(s)' showing examples of summative assessments they had recently used in a CLIL setting. They were provided with written instructions as to what to submit, including explanations of what was meant by terms such as 'assessment', 'summative', 'test' and 'assignment' in the context of this study, and asked to classify their materials as either tests or assignments based on those definitions (see Appendix S2). Requested for each bundle were:

1. All relevant documents pertaining to a summative assessment recently used in years 1-3 of BSE, i.e. the test or assignment instructions, assessment criteria/answer keys, and anything else the teacher considered relevant.
2. Three anonymized examples of completed and graded student products, preferably exemplifying strong, average and weaker performance.
3. A completed questionnaire (Appendix S2) regarding the teacher's background and teaching context, and information about the assessment materials.
4. A signed consent form.

In total, the thirteen participating teachers submitted seventeen assessment bundles. Two teachers submitted more than one bundle (T02_TEB: bundles A02-A05; T06_ST: bundles A09-A010). We decided to handle these as separate bundles, as they were unrelated assessments, for different classes, focusing on different content and skills, and displayed different features. It should be borne in mind, however, that the overlap in teachers may introduce some bias in the results. Table 1 shows how the bundles represented a range of subject areas, age-groups and assessment types, although written assessments predominate (14 out of 17). This is in line with the types of assessment teachers most often reported using in the questionnaire. With respect to educational stream, the over-representation of *vwo* (pre-university) assessments in the sample reflects the actual spread of Dutch BSE, which is most common in *vwo* tracks (Mearns et al., 2023). One assessment bundle (A11) did not contain graded examples of student work.

Table 1 Assessment bundles

Bundle	Teacher	Subject	Domain/topic	Test/ assignment	Assessment type	Year (Track)
A01	T01_TEB	English	Creative writing	Assignment	Written (diary entry)	2 (vwo)
A02	T02_TEB	English	Speaking	Assignment	Oral (short speech)	3 (havo)
A03			Grammar; writing	Test	Written (gap-fill tasks + application letter)	2 (havo)
A04			Creative writing	Assignment	Written (short story)	1 (vwo)
A05			Literature	Assignment	Written (tasks concerning a novel)	1 (vwo)
A06	T03_ST	Geography	Geology	Assignment	Oral (research presentation)	2 (vwo)
A07	T04_ST	History	World War I propaganda	Assignment	Written (research report)	3 (vwo)
A08	T05_ST	Geography	Wealth differences	Assignment	Written (research report)	1 (vwo)
A09	T06_ST	History	Napoleon	Assignment	Written (research report) or oral (research presentation) – student's choice	2 (vwo)
A10			Cold War	Test	Written (closed and open questions)	3 (vwo)
A11	T07_ST	Biology	Reproduction	Test	Written (closed and open questions)	2 (havo/ vwo)
A12	T08_ST	Biology	Senses, heredity and evolution	Test	Written (closed and open questions)	3 (vwo)
A13 (a/b)	T09_ST	Science	Light	Assignment	Written (research report; A13a) Product/design (spectroscope; A13b)	2 (vwo)
A14	T10_ST	Biology	Food and digestion	Test	Written (closed and open questions)	2 (vmbo/ havo/vwo)
A15	T11_ST	Maths	Statistics	Assignment	Written (research report)	2 (vwo)
A16	T12_ST	Chemistry	Classification of substances, reactions and energy	Test	Written (open questions)	3 (vwo)
A17	T13_ST	Biology	Plants	Assignment	Written (research report)	1 (havo/ vwo)

2.2.2.3 Analyses

Analysis of the assessment bundles and the accompanying background information provided by the teachers (Appendix S2) was carried out deductively, using a pre-determined set of themes formulated based on insights from research discussed above (Appendix S3). The first two themes coincide with those addressed in the questionnaire, namely: (1) the relative roles of content and language, and (2) the use of accommodations and translanguaging. Additionally, we analyzed (3) the cognitive and linguistic demands of each assessment.

Focus on content and/or language (theme 1) was identified from teachers' responses to the background questionnaire, and the weighting of and attention to content and language in grading and/or feedback, based on assessment criteria and graded student work. Regarding theme 2, we examined teachers' responses to the background questionnaire, and any relevant information in assessment instructions and tasks. Finally, to gather insight into the cognitive and linguistic demands (theme 3), we coded the assessments using a framework proposed by Lo and Lin (2014). Within this framework, CLIL assessments can be positioned in a three-by-three matrix, including cognitive (recall, application, analysis) and linguistic levels (vocabulary, sentence, text). At the recall level, students are asked to report or repeat what they have learned. For application, they apply what they have learned to new situations or problems. At the analysis level, they engage in higher-order thinking processes (e.g. synthesizing, evaluating). Along the linguistic dimension, assessment tasks can either require students to understand or produce language on word-level (e.g. subject-specific vocabulary), sentence-level (e.g. sentence patterns commonly used to explain, describe, etc.), and text-level (e.g. a subject-specific text genre).

Qualitative analysis was carried out by the second author, with samples dual-coded by the first author. Unresolved issues were discussed with the last author.

3 Results

We describe the results for each stage of data collection (questionnaire and assessment bundles) in turn.

3.1 Stage 1: Self-reported assessment practices (questionnaire)

3.1.1 *The relative roles of content and language in assessment*

Responses to the question "What is most commonly evaluated? (language, content, or both)" are summarized in Table 2. Most TEBs reported that they assess both language and content, either together (30 %) or in separate assessments (40 %). The majority of STs stated that they assess only content, although 19 % reported assessing both language and content, in most cases in combined assessments.

Table 2 Focus of assessments according to TEBs and STs

	TEBs (N = 10)		*STs (N = 32)*	
	n	**%**	**n**	**%**
Language only	2	20	0	0
Content only	0	0	26	81
Both (total)	8	80	6	19
– *Separately*	*4*	*40*	*2*	*6*
– *Together*	*3*	*30*	*4*	*13*
– *Not specified*	*1*	*10*	*0*	*0*

Table 3 Language aspects taken into account when grading tests and assignments

Language aspect in grade	**TEBs (N = 10)**				**STs (N = 32)**			
	Test		*Assignment*		*Test*		*Assignment*	
	n	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Pronunciation	3	30	5	50	0	0	3	9
Spelling	9	90	8	80	2	6	4	13
Vocabulary	9	90	9	90	5	16	4	13
Grammar	9	90	9	90	1	3	4	13
Structure	9	90	9	90	1	3	6	19
Clarity	6	60	10	100	5	16	7	22
Register	6	60	10	100	2	6	4	13
Subject-specific language	5	50	6	60	8	25	9	28

Teachers were asked which aspects of language they take into account when grading and when giving feedback. The results of these questions are summarized in Table 3 and Table 4, respectively.

With regard to both grading and feedback, TEBs again more often reported paying attention to language than did STs. The percentage of TEBs indicating that they paid attention to language aspects was between 50 % and 100 %, with the exception of the aspect of pronunciation in tests (30 %). For STs, the range was lower, namely for grading, between 0 % (pronunciation in tests) and 28 % (subject-specific language in assignments), and for feedback, between 3 % (pronunciation in tests) and 38 % (clarity and subject-specific language in assignments). It is worth noting that four STs who had first indicated that they evaluated only content, did indicate taking one or more language elements into account when grading. Apparently, when confronted with the

Table 4 Language aspects taken into account in feedback on tests and assignments

Language aspect in feedback	TEBs (N = 10)				STs (N = 32)			
	Test		*Assignment*		*Test*		*Assignment*	
	n	%	n	%	n	%	n	%
Pronunciation	3	30	8	80	1	3	8	25
Spelling	7	70	7	70	11	34	11	34
Vocabulary	8	80	9	90	8	25	10	31
Grammar	7	70	7	70	8	25	10	31
Structure	6	60	9	90	7	22	11	34
Clarity	6	60	9	90	10	31	12	38
Register	6	60	10	100	4	13	8	25
Subject-specific language	5	50	6	60	11	34	12	38

Table 5 Ranges of relative weighting of content, language, and other aspects when grading

Assessment type	Aspect	TEBs	STs
Tests	Language	50-100%	0-20%
	Content	0-20%	80-100%
	Other	0-10%*	0%
Assignments	Language	30-60%	0-30%
	Content	33-70%	60-100%
	Other	10-33%*	5-20%**

* rationale (test) & creativity (assignment); ** lay-out, structure, on-time, critical thinking (assignments)

precise terms for language aspects, more teachers realized they do take them into account.

In addition to which aspects were graded, teachers were also asked about the relative emphasis ('weighting') placed on content- and language-related aspects of assessment when grading. Table 5 shows the ranges of relative weighting as reported by TEBs and STs for aspects of language and content, for summative tests and assignments. In line with the previous findings, the ranges for TEBs lean more towards language and those for STs towards content. This difference is most pronounced in relation to tests.

Table 6 Accommodations allowed by TEBs and STs in tests and assignments

Accommodations	TEBs (N = 10)				STs (N = 31)			
	Test		*Assignment*		*Test*		*Assignment*	
	n	%	n	%	n	%	n	%
Dictionaries	5	50	9	90	6	19	16	52
Ask the teacher	2	20	7	70	13	42	22	71
Online translator	0	0	7	70	2	6	17	55
Glossaries	0	0	5	50	2	6	12	39
Personal idiom file	0	0	1	10	2	6	10	32
Personal notes	1	10	0	0	0	0	0	0
Ask classmates/parents	0	0	0	0	0	0	2	6
Pictures	0	0	1	10	10	32	10	32
Translations	0	0	0	0	7	23	9	29
Other: Dyslexia support	1	10	1	10	0	0	0	0

3.1.2 *Accommodations and translanguaging in assessment*

Table 6 displays the percentage of teachers reporting to allow different kinds of accommodations for tests and assignments, for TEBs and STs separately. Twenty-four teachers (out of a total of 41: one ST never used tests) reported to allow some kind of accommodations during test administration: asking the teacher and using dictionaries are most often allowed. For assignments (n = 41; one ST reported never to give assignments), most (34) teachers report that some type of accommodation is allowed.

Teachers appear to make little use of translanguaging strategies. Although up to 29 % of STs report providing some support for their assessments using translations, the main language of the instructions and of test and assignment formats was reported to be English. One ST adds that the choice of language(s) may depend on year. Likewise, student answers were also usually only allowed in English: 2 STs allowed some Dutch, one of whom added this was only allowed in year 1.

3.2 Stage 2: Actual assessment practices (assessment bundles)

We now move on to the results of the qualitative analysis of the assessment bundles submitted by teachers. For an overview of the contents of the bundles, see Table 1 (Method section). As with the questionnaire, the results will be presented under each of the three main analytical themes: roles of content and language, accommodations and translanguaging, and cognitive and linguistic demands.

3.2.1 *The relative roles of content and language in assessment*

Table 7 summarizes our findings regarding the relative roles of content and language in the assessments, based on three indicators: (1) the focus of each assessment as described in answer to the question, “What are the materials aimed at assessing? What did you want students to show they knew/could do?”; (2) the relative weighting of content and language in grading, as evidenced in assessment criteria and assigned grades; and (3) the focus of feedback provided by teachers in sample learner products.

The patterns visible in Table 7 resemble those observed in the questionnaire data. The reported role of language is greater in TEBs’ assessments than in STs’ assessments, whereas for content we see the opposite. In describing the focus of assessments, TEBs mostly mention linguistic knowledge and skills, with the exception of ‘*understanding of literature*’ as an aim of A05. STs, on the other hand, describe the focus of their assessments almost exclusively in terms of content knowledge and skills, with no explicit mention of language. Some of the descriptions do refer to linguistic aspects implicitly, by mentioning subject-relevant language functions (e.g., ‘*explain*’; A08) or genres (e.g., ‘*writing a lab report*’; A17).

When looking at the weighting of different aspects in assessment criteria, a similar pattern emerges. In TEBs’ assessments, the weighting of language is 33%-100%, and that of content 10%-30%. In STs’ assessments, content weighs 40%-100%, whereas the weight of language is 0%-20%. Other assessment criteria include creativity, text structure or organization, lay-out, referencing, active participation, and presentation skills, and weigh 0%-40% in TEBs’, and 0%-53% in STs’ assessments.

To provide more insight into the role of content in TEBs’ grading and of language in STs’ grading, Table 7 also specifies which content-related (for TEBs) and linguistic (for STs) aspects were actually present in assessment criteria. From the four TEB assessments in which content plays a role in grading (A01, A02, A04, A05), one assessment (A05) clearly focuses explicitly on subject-relevant content (literature). In one other case (A02), the content assessed does not represent cultural or linguistic knowledge that could be considered part of the English-as-discipline curriculum (e.g. ‘*appear*[*ing*] *knowledgeable*’ on any chosen presentation topic). In the remaining two cases, the content aspects assessed relate to students’ creative writing skills, namely, ‘*the use of supporting details*’ in a diary entry (A01) and the ‘*plot, character, idea/message*’ of a short story (A04). When language plays an explicit role in STs’ grading (A06, A08, A09, A13, A15, A17), the focus is usually on accuracy of formulations (word choice, grammar, spelling, punctuation). In two cases (A06, A15), it remains unclear which linguistic aspects are taken into account as the relevant criteria are broadly formulated as ‘*use of English*’ or ‘*use of language*’. In one case (A09), students are assessed on using ‘*their own words*’.

Finally, from analysing teachers’ feedback on student work the first observation that stands out is the absence of feedback in ten of the assessment bundles. In the seven cases where feedback was provided, four assessment bundles included feedback on both content and language (3 times by the same TEB: A02, A04, A05, once by ST: A15),

Table 7 Relative roles of content and language in assessment bundles

Assessment and teacher	**Subject, domain/topic**	**(1) Focus/goal of assessment (as described by teachers)**	**Role of content and language (as apparent in assessment criteria and graded student work)**	
	Assessment type		**(2) *Weighting***	**(3) *Feedback***
A01; T01_TEB	English; creative writing Diary entry	Creativity, spelling/grammar, word choice, fluidity, organization	Language: 60 % *Content:* 20 % (development: use of supporting details) *Other:* 20 % (creativity)	No feedback (other than completed rubric)
A02; T02_TEB	English; speaking Speech	Mostly speaking skills, also critical thinking, presenting/organizing information	Language: 60 % *Content:* 10 % (subject knowledge: was the subject appropriate for the topic/did speaker appear knowledgeable) *Other:* 30 % (presenting skills)	Feedback on both language and content
A03; T02_TEB	English; grammar/ writing Gap-fill tasks + application letter	Mostly grammar, also vocabulary, writing skills	Language: 100 %	Feedback on language
A04; T02_TEB	English; creative writing Short story	Mostly creative writing skills; also grammar, vocabulary, text organization	Language: 70 % *Content:* 30 % (plot, character, idea/message)	Feedback on both language and content
A05; T02_TEB	English; literature Booklet with tasks concerning a novel	Mostly understanding of novel; also writing, executive skills (organizing work, keeping up with homework, etc.)	Language: 33 % *Content:* 27 % (accurate completion of tasks) *Other:* 40 % (creativity; active participation)	Feedback on both language and content
A06; T03_ST	Geography; geology Research presentation	General system of plate tectonics; specifics on one kind of natural disaster	Content: 40 % *Language:* 7 % (use of language) *Other:* 53 % (presentation skills)	No feedback (other than completed rubric)
A07; T04_ST	History; World War I propaganda Research report	Knowledge of propaganda techniques, analyzing propaganda, creating propaganda	Content: 100 % Language: 0 %	Feedback on content

Table 7 Relative roles of content and language in assessment bundles (*cont.*)

Assessment and teacher	**Subject, domain/topic**	**(1) Focus/goal of assessment (as described by teachers)**	**Role of content and language (as apparent in assessment criteria and graded student work)**	
	Assessment type		**(2) *Weighting***	**(3) *Feedback***
A08; T05_ST	Geography; wealth differences Research report	Identifying indicators used to measure development; ranking countries based on data; analysing research data; explaining ranking of countries	Content: 60 % *Language:* 20 % (sentence fluency) *Other:* 20 % (design, lay-out, organization)	No feedback (other than completed rubric)
A09; T06_ST	History; Napoleon Research presentation or report	Research, presentation	Content: 57 % *Language:* 14 % (use of own words) *Other:* 20 % (structure, lay-out, references)	No feedback (other than completed rubric)
A10; T06_ST	History; Cold War Written test with closed and open questions	(Use of) knowledge	Content: 100 % Language: 0 %	No feedback
A11; T07_ST	Biology; reproduction Written test with closed and open questions	Learning and understanding skills	Content: 100 % Language: 0 %	Unknown (no student work available)
A12; T08_ST	Biology; senses, heredity and evolution Written test with closed and open questions	Names of parts of the senses and their functions, how evolution works (describe in their own words), how Mendelian genetics works, making Punnet squares and calculations	Content: 100 % Language: 0 %	No feedback
A13 (a & b); T09_ST	Science; light Research report and product/design (spectroscope)	Writing a scientific report; following instructions for experiment of their choice; explaining phenomena related to light that occurred with experiment	Content: 86% *Language:* 7 % (accurate use of (scientific) language) *Other:* 7 % (lay-out)	No feedback (other than completed rubric)

Table 7 Relative roles of content and language in assessment bundles (*cont.*)

Assessment and teacher	**Subject, domain/topic**	**(1) Focus/goal of assessment (as described by teachers)**	**Role of content and language (as apparent in assessment criteria and graded student work)**	
	Assessment type		**(2) *Weighting***	**(3) *Feedback***
A14; T10_ST	Biology; food and digestion Written test with closed and open questions	Theory	Content: 100% Language: 0%	On content; one instance of feedback on language ('strange sentence')
A15; T11_ST	Maths; statistics Research report	Basic concepts of statistics: calculating mean, median, mode; being able to make several kinds of charts	Content: 90% *Language:* 10% (use of English)	Feedback on content and language (i.e., grammar, spelling, punctuation)
A16; T12_ST	Chemistry; classification of substances, reactions and energy Written test with open questions	Recognising an oxide from its chemical symbol, writing balanced chemical equations, understanding of the different types of chemical bonding using Bohr diagrams	Content: 100% Language: 0%	No feedback
A17; T13_ST	Biology; plants Research report	Research skills: designing an experimental set-up, executing a research plan, collecting data, making tables/graphs, reflecting on process, writing a lab report	Content: 93% *Language:* 7% (grammar, spelling, punctuation)	No feedback (other than completed rubric)

in two instances the feedback only related to content-related aspects (both by STs: A07, A14), and in the final case, only feedback on language was provided (by TEB: A03).

In most cases, what teachers describe to be the focus of their assessments, the assessment criteria, and the foci in their feedback, are aligned. In some bundles, however, this is not the case. Firstly, most STs who take linguistic aspects into account when grading do not make explicit reference to those aspects in their description of assessment goals (Table 7, column 3). Secondly, in the two cases where there is more implicit mention of linguistic aspects such as subject-specific language functions or genres in STs' assessment descriptions (i.e., A13, A17), only in one case (A13) do the assessment criteria also foreground the use of subject-specific language (although formulated quite broadly:

'*accurate use of* (*scientific*) *language*'). In A17 on the other hand, while the focus in the description is on a subject-specific genre ('*writing a lab-report*'), the assessment criteria concern accurate use of grammar, spelling, and punctuation. Finally, in one case (A14), the ST provided feedback on language (one instance only) whereas language was not a focus in grading or assessment criteria.

3.2.2 *Accommodations and translanguaging in assessment*

In the background questionnaire, we asked teachers which accommodations they allowed their students to use during assessment. We also analysed assessment instructions to determine how teachers used accommodations and translanguaging. Moreover, assessment tasks as well as student products were examined to establish which languages were used in assessments.

With respect to accommodations, a clear difference emerged between tests and assignments. While no accommodations were apparently incorporated or allowed during tests, a variety of support materials were available during completion of assignments: namely, online sources/internet (5×); any type of resource (3×); books (3×); (online) dictionaries (2×); lesson materials, such as PowerPoints or instructional videos (2×); cue cards (1×); ask the teacher (1×). Conversely, no examples were found of accommodations incorporated in assignment format, such as supporting visuals, graphic organisers or writing frames.

Regarding the use of translanguaging, a similarly monolingual image arises from the assessment bundles as from the questionnaire data. In fact, while a small number of questionnaire respondents indicated that they provided some support in Dutch, only English was used in the actual assessment instructions and formats submitted for analysis. Likewise, in the assessment bundles, students were only allowed to use English to complete the tasks.

3.2.3 *Cognitive and linguistic demands of assessments*

To map the cognitive and linguistic demands of the assignments, we used the framework proposed by Lo and Lin (2014), described in the Method section. Figure 1 illustrates where the assessments can be positioned in the matrix. Whereas Lo and Lin differentiate between assessments that require comprehension only (i.e. receptive level) and assessments in which students need to produce language (i.e. productive level), we chose not to make this distinction in Figure 1, as all assessments analysed necessitated language reception (e.g. reading assessment task, consulting sources) as well as production (e.g. answering open test questions, writing a research report). Moreover, assessments often pose demands on different levels (e.g. content analysis also implies content recall; production at text level inherently requires production at sentence level). We therefore chose to classify assessments based on the highest (cognitive and linguistic) level they require.

Three of the five assessments submitted by TEBs (A01, A03, A04) could not be placed in the matrix. Assessment A03, a fully language-focused test, could only be scored on

Figure 1 Cognitive and linguistic demands of assessments, based on Lo and Lin (2014)

Linguistic demands	Cognitive demands		
	Recall	Application	Analysis
Vocabulary			

Sentence		A11 (havo/vwo; 2);
		A12 (vwo; 3);
		A14 (vmbo/havo/vwo; 2)
			A10 (vwo; 3);
			A16 (vwo; 3)

Text			A02 (havo; 3);
			A05 (vwo; 1);
			A06 (vwo; 2);
			A07 (vwo; 3);
			A08 (vwo; 1);
			A09 (vwo; 2);
			A13 (vwo; 2);
			A15 (vwo; 2);
			A17 (havo/vwo; 1)

the linguistic dimension, and therefore does not appear in Figure 1. In A01 and A04, the aspects identified by teachers to represent content relate to creative writing (e.g. plot, character details). From the data provided in the assessment bundles, it was not clear to what extent these aspects followed explicit instruction on literary competence, or whether they were simply vehicles for learners to demonstrate their writing skills. This contrasts with the role of creative writing assignments in assessment A05, which build on exploration of the techniques used in a literary work. Furthermore, the lack of detail in the materials (which included learner work and completed rubrics, but no written instructions) prevented us from being able to gauge the cognitive demands of the tasks assigned. These two assessments therefore also do not appear in Figure 1.

As Figure 1 demonstrates, both cognitive and linguistic demands of assessments are generally high. None of the assessments remain at the lower demand recall and/or vocabulary levels. Three assessments could be characterized to pose medium cognitive and linguistic demands (e.g. a written test with open questions in which students have to apply learned content, and formulate answers using full sentences). Two assessments likewise demand language production at the sentence level, but require cognitive processes at the analysis level (e.g. a written test with open questions which require students to perform an evaluation or comparison). The majority of assessments (9 out of the 14 assessments of which both the cognitive and linguistic demands could be analysed), however, can be positioned in the most demanding cell of the matrix. In those assessments, students are expected to perform complex cognitive processes and report on them in written or oral texts (e.g. conducting and reporting on an experiment).

When we examine the years and educational tracks in which the different assessments were used, it does not seem to be the case that cognitive and/or linguistic demands increase with school year (1-3). In fact, the three year 1 assessments (A05, A08, A17) can all be characterized as maximally demanding, both cognitively and linguistically, whereas three out of the five year 3 assessments are less demanding (A10; A12, A16). Similarly, no clear pattern emerges with respect to the different educational tracks (pre-vocational education: *vmbo*, general higher education: *havo*; pre-university education: *vwo*). Both *havo* and *vwo* assessments can be found in the cell representing the lowest demands in our sample (application/sentence) as well as in the most demanding category (analysis/text). Since there is only one assessment which is (also) used in the *vmbo* track, no meaningful conclusion can be drawn about the demands of *vmbo* assessments as compared to assessments for *havo* and *vwo*.

4 Discussion

This small-scale study aimed to increase our understanding of ways in which teachers of English in bilingual education (TEBs) and CLIL subject teachers (STs) in a Dutch CLIL context approach summative assessment. Two types of data collection were used: a self-report questionnaire to take inventory of perceived practices, and in-depth analysis of assessment materials to gain insight into actual (or operational) practices (van den Akker, 2003). The assessments submitted were largely written tests and assignments, in line with findings from previous CLIL research highlighting the predominance of written assessment (Otto, 2017; Reierstam & Sylvén, 2019). Together, the two data sources shed light on: (1) the interplay between content and language in these assessments; (2) the use of assessment accommodations and translanguaging as means to mitigate potential interference of lower L2 proficiency; and (3) the cognitive and linguistic demands assessments pose on students. Below, we discuss the significance of the findings in each of these areas, in relation to the limited literature on summative assessment practices in CLIL and other L2 content teaching settings.

4.1 Relative roles of content and language

It is perhaps not surprising that TEBs placed more emphasis on language in assessment, and STs on content. This trend was apparent in both the self-report data and in the grading and feedback practices represented in the assessment bundles. This reflects earlier findings suggesting that teachers' disciplinary identity influences the focus of their assessments (Otto & Estrada, 2019).

The content observed in TEBs' assessment materials was not always linked to subject-relevant learning objectives (e.g. belonging to the fields of literature, cultural studies or linguistics), but in some cases was used as a vehicle to assess a linguistic goal (e.g. 'appearing knowledgeable' when giving a speech). A complexity that we observed when analysing the materials provided by teachers was the position of creative writing. Teachers identified content criteria in creative writing exercises, although in two of the three examples of this, it was not clear whether and how those criteria aligned with prior learning related to literary competence, or if they were extensions of learners' writing skills. Literature on CLIL rarely refers to creative writing, and when it does, it tends to be positioned as a means of processing input (e.g. exploring a scene from a play, as in Dale (2020)) or of practising writing skills (e.g. Weiss et al., 2023), rather than of demonstrating literary competence as a form of content. That teachers in this study appeared to consider creative writing to have a content goal, although that goal was not always aligned with our understanding of disciplinary content, highlights the lack of clarity regarding what constitutes disciplinary content in the language classroom.

When STs assessed language, they appeared to pay most attention to lower-order aspects of language (e.g. spelling, grammar, punctuation, pronunciation) or to unspecified language "use", in spite of assessments being linguistically and cognitively demanding. While self-report data suggested that subject-specific language use was sometimes a focus of grading and feedback, this was rarely reflected in the assessment bundles. This echoes earlier findings from the Dutch CLIL context (Busz et al., 2014), where teachers provided micro-level feedback similar to that found in the current study, as opposed to feedback on subject-specific communication. Our findings might suggest that the increasing emphasis on disciplinary literacies in CLIL (Coyle & Meyer, 2021) as key to subject learning is not yet reflected in teaching practice, at least in terms of assessment. As Lo et al. (2019) emphasise, this could limit CLIL's effectiveness in terms of the development of subject-specific language, as washback from assessment could lead STs to neglect the linguistic aspects of their subject in their teaching. Furthermore, in line with earlier findings by Hönig (2010), analysis of the assessment materials and self-report data suggest that language does play a role in STs' assessment of content, even when it is not identified as a goal of assessment or mentioned explicitly in instructions. As identified by Lo and Fung (2020), this could lead to underestimation of learners' content knowledge or to learners being penalized on the basis of "invisible" language criteria (Llinares et al., 2012, p. 274).

4.2 Accommodations and translanguaging

CLIL teachers in this study reported that they allow room for a range of accommodations to support completion of summative assignments (e.g. projects, presentations), and to a lesser extent for timed tests taken in classroom settings. In the assessment bundles, we saw that learners were allowed to draw on diverse support materials when working on summative assignments, although we found few instances of embedded accommodations such as visual support or text modification, to help learners understand input or instructions, or graphic organisers to support output. This contrasts to findings from a similar exploration by Otto (2017), who found visual support and guided questioning to be the most common forms of accommodation in the CLIL assessments she analysed, although she too observed an absence of support for language production. As earlier studies have suggested that a variety of accommodation strategies is needed to respond to different learner needs (De Backer et al., 2019a), this appears to be an area in which CLIL teachers may be able to support their learners better. As the multiple-choice options in the questionnaire were limited, however, we do not know whether this trend extends beyond the seventeen assessments analysed. Further research could explore this aspect of CLIL assessment in more detail, in order to make more concrete recommendations.

Translanguaging in assessment appeared not to be common practice. This confirms our expectation based on the known tendency for languages other than English to be excluded from CLIL classrooms in the Netherlands (van Kampen et al., 2018). On the few occasions where translanguaging was used to support learners or permitted in learners' responses, teachers qualified their responses by stating that this was permitted only in the first year. This may reflect Oattes et al.'s (2018) observation that STs are implicitly aware that translanguaging can be a useful pedagogical tool, and tend to use it mostly with younger learners. In light of Oattes et al.'s (2020) finding that assessing first year CLIL learners' content knowledge in English can be detrimental to their performance, this might be considered a well-reasoned approach.

4.3 Cognitive and linguistic demands

As exemplified in the analysed assessment bundles, both cognitive and linguistic demands of content-focused assessment were generally high. This was the case irrespective of year (1, 2 or 3) and educational track (pre-vocational/*vmbo*, general/*havo* or pre-university/*vwo*). Lo and Fung (2020) warn about the risks this combination of high demands can pose to assessment validity due to learners' developing L2 proficiency, as they found that increased linguistic demands adversely affected learners' performance in terms of content. In the context of our study where, in contrast to Lo and Fung's findings, high demands on both axes were found in the earliest years of CLIL, and in light of Oattes et al.'s (2020) conclusion that language can be a barrier to younger learners in CLIL content assessments, the question can be posed as to whether it is fair and valid to assess them in this way. That said, emphasis on disciplinary literacies as integral to content learning emphasizes the inextricable connection between language and cognition in any learning situation (Dalton-Puffer, 2013). From this perspective, higher cognitive demands and higher linguistic demands will likely go hand in hand, but need to be built-up in structured and scaffolded ways (Coyle & Meyer, 2021). Rather than suggesting that teachers lower demands in assessments, it might be better to address the use of scaffolded approaches to teaching and assessment (e.g. using accommodations and translanguaging), in CLIL teacher education and professional development.

5 Conclusions and implications

This study was limited by its relatively small number of respondents. Nonetheless, the combination of exploring practice on perceived and operational levels allowed for in-depth analysis and triangulation, in order to paint a preliminary picture of current CLIL assessment practices in bilingual lower-secondary education in the Netherlands. While its findings cannot be generalised to the whole population of CLIL teachers in the Netherlands, we hope this study will be a springboard for further research, for example examining the alignment between classroom teaching and assessment, or exploring assessment programmes as a whole, to identify whether the distribution of attention for language and content balances out over the course of an academic year or more. In-depth interviews with teachers could be a valuable supplement to these approaches, in order to gain insight into the thinking behind assessment choices and in what teachers need in order to develop their practice further in this regard. In so doing, more emphasis could also be placed on exploring the relative roles and perspectives of language teachers and teachers of other subjects. As previous research has suggested that experience with CLIL may influence teachers' practice in mainstream classes (Oattes et al., 2018), and considering the similarities between CLIL and teaching practices in other multilingual classrooms (Hajer, 2018), all of the areas mentioned above could be explored in relation to CLIL specifically, or in other multilingual or monolingual settings.

Returning to Sato's (2023) model of 'separate', 'weakly integrated' and 'strongly integrated' approaches to CLIL assessment, the findings of this study suggest that, in practice, the lines between these approaches are blurred. As has also been found elsewhere, while STs appeared to believe that they handled language and content separately, the intrinsic relationship between linguistic and content elements nonetheless appeared to play a role in assessment. Learners were often allowed access to support materials, which might contribute to mitigating interference of language proficiency in content assessment. On the whole, however, little evidence was found of adaptations in the format or design of assessments in order to directly support understanding or production of language, nor of opportunities to use translanguaging for this purpose, in spite of both cognitive and linguistic demands of assessments being high. English teachers, conversely, did pay explicit attention to language, but not in relation to meaningful content. This could be a missed opportunity to support learners' development in the literacies of English as a discipline in its own right, in particular with an eye to the requirements of the content-heavy IB programme in the senior years.

As the goal of CLIL is to support learning of both content and language in an integrated way, we would argue for also integrating and meaningfully addressing these elements in assessment, in ways that go beyond evaluating general "language use" in content subjects and "appearing knowledgeable" about content in language subjects. This can be achieved through high-demand, high-support approaches that acknowledge both content and language as integral to subject learning, while also acknowledging that language proficiency can affect communication of content knowledge, and therefore providing appropriate linguistic support. As emphasised in work on disciplinary literacies (e.g. Coyle & Meyer, 2020), similar issues are at stake even when learning takes place in the learners' L1, as subject learning inevitably involves developing proficiency in disciplinary discourse. Furthermore, addressing disciplinary content in L1 or any foreign language classrooms could help expose more learners to the benefits of content and language integration (Mearns & Platteel, 2020; Michel et al., 2021). To this end, teachers in any setting could benefit from support in developing integrated assessment practices, with appropriate levels of support, through focused attention as part of teacher education and professional development, from a multilingual disciplinary literacies perspective. Coupling professional development activities such as these to teacher action research or professional learning communities could serve the dual purpose of supporting teachers' development and providing valuable new insights for the broader CLIL and disciplinary literacies communities.

Author contributions

Tessa Mearns: conceptualisation, methodology, investigation, validation, writing – original draft, writing – reviewing and editing; Catherine van Beuningen: conceptualisation, methodology, investigation, formal analysis, writing – original draft, writing – reviewing and editing; Niels Nederlof: methodology, investigation, data curation; Nivja H. de Jong: conceptualisation, methodology, investigation, formal analysis, writing – reviewing and editing.

Funding information

Funding for this research was provided by the Meesterschapsteam Moderne Vreemde Talen, part of the programme Meesterschap in de Vakdidactiek from Vakdidactiek Geesteswetenschappen.

Statement of interest

The authors have no known competing financial interests or personal relationships that could have influenced the work reported in this manuscript.

Statement of technology use

No AI-based generative technology was used in the preparation of this manuscript and the execution of the research that the manuscript reports upon.

Supporting information

In the supplementary file:

- Appendix S1: Teacher questionnaire on self-reported assessment practices
- Appendix S2: Background questionnaire on assessment bundles
- Appendix S3: Coding Scheme for Assessment Bundles

References

Bentley, K. (2010). *The TKT Course CLIL Module*. Cambridge University Press.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*, 347-364. https://doi.org/10.1007/BF00138871

Busz, M., Helleman, J., DeVincent, D., Verwoerd-Sowariraj, S., & Tonsberg Schlie, K. (2014). De praktijk van feedback op Engelstalige profielwerkstukken [The practice of giving feedback on extended essays written in English]. *Levende Talen Tijdschrift*, *15*(4), 26-37. https://lt-tijdschriften.nl/ojs/index.php/ltt/article/view/903

Butler, F.A., & Stevens, R. (1997). *Accommodation strategies for english language learners on large-scale assessments: Student characteristics and other considerations.*

Cenoz, J. (2017). Translanguaging in school context. International perspectives. *Journal of Language, Identity and Education*, *16*(4), 193-198. https://doi.org/https://doi.org/10.1080/15348458.2017.1327816

Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and Language Integrated Learning*. Cambridge University Press.

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

Dale, L. (2020). *On Language Teachers and CLIL. Shifting the Perspectives* University of Amsterdam]. Amsterdam. https://dare.uva.nl/search?identifier=98c15569-9e2f-4542-8472-9515aa7bac7d

Dale, L., Oostdam, R., & Verspoor, M. (2018). Juggling ideals and constraints. The position of English teachers in CLIL contexts. *Dutch Journal of Applied Linguistics*, *7*(2), 177-202. https://doi.org/doi.org/10.1075/dujal.18002.dal

Dale, L., & Tanner, R. (2012). *CLIL Activities: A Resource for Subject and Language Teachers*. Cambridge University Press.

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

Dalton-Puffer, C., Llinares, A., Lorenzo, F., & Nikula, T. (2014). "You Can Stand Under My

Umbrella": Immersion, CLIL and Bilingual Education. A Response to Cenoz, Genesee & Gorter (2013). *Applied Linguistics, 35*(2), 213-218. https://doi.org/https://doi.org/10.1093/applin/amu010

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/doi:10.1515/eujal-2017-0028

De Backer, F. (2020). Bridging the gap between learning and evaluation. Lessons learnt from multilingual pupils. *Journal of Applied Linguistics and Professional Practice, 14*(1), 96-118. https://doi.org/https://doi.org/10.1558/jalpp.39770

De Backer, F., Baele, J., van Avermaet, P., & Slembrouck, S. (2019a). Pupils' Perceptions on Accommodations in Multilingual Assessment of Science. *Language Assessment Quarterly, 16*(4-5), 426-446. https://doi.org/https://doi.org/10.1080/15434303.2019.1666847

De Backer, F., Slembrouck, S., & van Avermaet, P. (2019b). Assessment accommodations for multilingual learners: pupils' perceptions of fairness. *Journal of Multilingual and Multicultural Development, 40*(9). https://doi.org/https://doi.org/10.1080/01434632.2019.1571596

Hajer, M. (2018). Teaching content through Dutch as a second language. How 'Language Oriented Content Teaching' unfolded in mainstream secondary education. *Dutch Journal of Applied Linguistics, 7*(2), 246-263. https://doi.org/10.1075/dujal.18001.haj

Hönig, I. (2010). Assessment in CLIL – A case study. *Vienna English Working Papers: Current Research on CLIL, 19*(3).

Llinares, A., Morton, T., & Whittaker, R. (2012). *The Roles of Language in CLIL.* Cambridge University Press.

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. https://doi.org/https://doi.org/10.1080/13670050.2018.1436519

Lo, Y.Y., & Lin, A.M.Y. (2014). Designing assessment tasks with language awareness: Balancing cognitive and linguistic demands. *Assessment and Learning, 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. https://doi.org/https://doi.org/10.1075/jicb.18028.lo

Mearns, T., & de Graaff, R. (2018). Bilingual education and CLIL in the Netherlands. The paradigm and the pedagogy. *Dutch Journal of Applied Linguistics, 7*(2), 122-128. https://doi.org/10.1075/dujal.00002.int

Mearns, T., & Platteel, T. (2020). Exploring teacher support for a content and language integrated modern languages curriculum. *Language, Culture and Curriculum*, 1-17. https://doi.org/10.1080/07908318.2020.1809665

Mearns, T., van Kampen, E., & Admiraal, W. (2023). CLIL in The Netherlands: Three decades and innovation and development. In D.L. Banegas & S. Zappa-Hollman (Eds.), *The Routledge Handbook of Content and Language Integrated Learning* (pp. 403-418). Routledge.

Mehisto, P., & Ting, T. (2017). *CLIL essentials for secondary school teachers*. Cambridge University Press.

Michel, M., Vidon, C., De Graaff, R., & Lowie, W. (2021). Language Learning beyond English in the Netherlands: A fragile future. *European Journal of Applied Liguistics*, *9*(1), 159-182. https://doi.org/https://doi.org/10.1515/eujal-2020-0020

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/https://doi.org/10.5565/rev/clil.33

Nuffic. (2019). *Kwaliteitsstandaard tweetalig onderwijs 2.0* [Quality Standard for Bilingual Education 2.0]. The Hague: Nuffic Retrieved from https://www.nuffic.nl/publicaties/kwaliteitsstandaard-tweetalig-onderwijs-20/

Oattes, H., Fukkink, R., Oostdam, R., de Graaff, R., & Wilschut, A. (2020). A showdown between bilingual and mainstream education: the impact of language of instruction on learning subject content knowledge. *International Journal of Bilingual Education and Bilingualism*, 1-14. https://doi.org/10.1080/13670050.2020.1718592

Oattes, H., Oostdam, R., de Graaff, R., & Wilschut, A. (2018). The challenge of balancing content and language: Perceptions of Dutch bilingual education history teachers. *Teaching and Teacher Education*, *70*, 165-174. https://doi.org/10.1016/j.tate.2017.11.022

Otto, A. (2017). *Assessment in CLIL: The Balance Between the Content and the Language. Madrid Bilingual Secondary Schools as a Case Study* [Unpublished, University of Alcalá]. Madrid.

Otto, A. (2019). 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). https://doi.org/10.5294/laclil.2018.11.2.6

Otto, A., & Estrada, J.L. (2019). Towards an Understanding of CLIL 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/https://doi.org/10.5565/rev/clil.11

Reierstam, H., & Sylvén, L.K. (2019). Assessment in CLIL. In L.K. Sylvén (Ed.), *Investigating Content and Language Integrated Learning: Insights from Swedish High Schools*. Multilingual Matters.

Sato, T. (2023). Assessment in CLIL. In D.L. Banegas & S. Zappa-Hollman (Eds.), *The Routledge Handbook of Content and Language Integrated Learning*. Routledge.

Serra, C. (2007). Assessing CLIL at Primary School: A Longitudinal Study. *International Journal of Bilingual Education and Bilingualism*, *10*(5), 582-602. https://doi.org/https://doi.org/10.2167/beb461.0

Sluijsmans, D., Joosten-ten Brinke, D., & van der Vleuten, C. (2013). *Toetsen met leerwaarde; Een reviewstudie naar de effectieve kenmerken van formatief toetsen* [Testing with value for learning; A review study on characteristics of effective formative assessment].

Trenkic, D., & Warmington, M. (2019). Language and literacy skills of home and international university students: How different are they, and does it matter? *Bilingualism: Language and Cognition*, *22*(2), 349-365. https://doi.org/https://doi.org/10.1017/S136672891700075X van Beuningen, C., & Polišenská, D. (2019). Meertaligheid in het voortgezet onderwijs: Een inventarisatiestudie naar opvattingen en praktijken van talendocenten [Multilingualism in secondary education: An inventory study into the views and practices of language teachers]. *Levende Talen Tijdschrift*, *20*(4), 25-36.

van den Akker, J. (2003). Curriculum perspectives: An introduction. In J. van den Akker, W. Kuiper, & U. Hameyer (Eds.), *Curriculum landscapes and trends* (pp. 1-10). Kluwer. https://doi.org/https://doi.org/10.1007/978-94-017-1205-7

van Kampen, E., Mearns, T., Meirink, J., Admiraal, W., & Berry, A. (2018). How do we measure up? A review of Dutch CLIL subject pedagogies against an international backdrop. *Dutch Journal of Applied Linguistics*, *7*(2), 129-155. https://doi.org/10.1075/dujal.18004.kam

Weiss, D., Gierlinger, E.M., & Hütter, J. (2023). Teaching genre-based writing in CLIL secondary education. The implementation of a language learning model in the subject Pedagogy. In E.M. Gierlinger, M. Döll, & G. Keplinger (Eds.), *TALK in Multilingual Classrooms* (pp. 261-296). Waxmann.

Yang, X. (2020). Assessment accommodations for emergent bilinguals in mainstream classroom assessments: a targeted literature review. *International Multilingual Research Journal*, *14*(3), 217-232. https://doi.org/10.1080/19313152.2019.1681615