Towards improved teacher CLIL practice.

Identification of design principles for a rubric styled feedback tool on teacher CLIL competencies

By:
Danny Desar
Bianca van der Molen
Tomas van der Vinne

Utrecht University
U-TEAch
Supervisor: F.A.N. Kranenburg
Date: 27-08-2012

Content

Abstract

Teachers in TTO and international schools teach both content and language. To do this effectively they use a method called "Content and Language Integrated Learning" (CLIL). In this research we identify design principles for creating a rubric-styled feedback tool for effective CLIL practice. We distil the main criteria for effective CLIL practice from the academic literature and validate it using an expert group of TTO teachers. Using descriptors of our own design, we generate feedback on the design principles for such descriptors using the same expert group of TTO Teachers.

Approximately sixty-six percent of the criteria from the literature is validated by our expert group. However, it is noted that the degree of validation is dependent on the personal experience and familiarity of different teachers with the various aspects of CLIL.

We show that descriptors must be very specific with no room for interpretation or discussion. Furthermore, they must describe realistic levels of competence and should avoid the quantification of things such as percentage of participation and attention levels, which are found to be immeasurable. The results of this research provide a solid basis for the creation of a rubric-styled feedback tool for the effective implementation of CLIL techniques.

Introduction

Problem statement

The school where this research took place wanted to analyse and improve the "Content and Language Integrated Learning" (CLIL) skills of their teachers. CLIL is used mainly at TTO and international schools to teach learners both subject related content as well as the target language. The head of the TTO department had expressed his concern for a lack of constructive feedback options for his TTO teachers to work on their CLIL competencies. The head of the school feared that the lack of feedback on CLIL practice might affect future improvement of TTO education at the school. His concern seemed justified as constructive feedback is generally known to be one of the main aspects for the improvement of a teacher's teaching practice. Having a tool that can provide such feedback on CLIL skills thus seems essential for improving good CLIL practice for TTO teachers. A feedback tool for CLIL does not exist yet and in this research we made the first steps towards creating such a tool. This research identifies design principles for the creation of a rubric styled feedback tool for CLIL and thus provides a stepping stone to improved teacher CLIL practice.

For this research we had two meetings with a group of TTO teachers from the research school. These teachers are our respondents and are referred to as the: "Expert group of TTO teachers". All the members of this expert group of TTO teachers taught different subjects and at different levels. The meetings were used to validate the literature criteria for good CLIL practice. Also, it gave us insight in the key aspects of workable descriptors. The acquired knowledge was used to make a first step in the design (criteria only) of six rubrics, one for each CLIL competency. The identification of design principles for the creation of a feedback tool for CLIL practice brought the creation of a completely operational feedback tool one step closer. This makes the findings of this research not only relevant for the research school but for all Dutch schools providing TTO education.

Theoretical Framework

The focus of this research is to identify design principles for the creation of a feedback tool for effective CLIL practice. An assessment of the available academic literature on feedback tools for CLIL practice showed this specific topic was not researched yet.
Therefore, we shifted our focus to finding academic literature that could help us successfully identify design principles for a feedback tool for CLIL. We found that useful research had been conducted with regards to the use of rubrics as an effective format for a feedback tool (Levi, 2005). Also, we learned how criteria that one might want to use in a rubric styled feedback tool can be successfully validated (McNamara, 1996; Baartman, 2008). Studies like these are very valuable to this research as they provide the tools we needed to identify and validate a set of design principles for a feedback tool for effective CLIL practice.

The use of CLIL in Dutch TTO schools

In modern Dutch TTO schools, CLIL is becoming an increasingly important aspect of the teaching practice. This is not surprising as the effective use of CLIL skills has great potential. It can lead to effective teaching of content and language, rather than simply teaching a subject in a foreign language (Dale, 2010). However, teaching from a CLIL perspective requires a change in overall teaching strategy. It requires educational materials that put a clear focus on the role of language in the learners' understanding of content related concepts (Marenzi et al., 2010)
When using CLIL, teachers should focus on scaffolding the students' knowledge and enhancing learner motivation and understanding. Furthermore, the students have to be involved in their own learning process via activating teaching strategies. This promotes the production of content and language specific output which in turn can be assessed with learning as its main aim (Dale, 2010).

As with any teaching strategy, the teacher plays a vital role in its success and good CLIL practice is no exception. This statement is supported by the authors of 'CLIL Skills' by Dale (2010) who state:

''As many CLIL teachers will testify, teaching a school subject through a second language brings with it a variety of challenges. How can subject teachers make sure that learners understand everything they need to know about the subject when a second language is being used by both the teacher and the learners? How can teachers' help learners acquire not only the content of their subject but also the language they need to demonstrate their understanding of the content? How can learners learn both content and language at the same time?''

Hence, capable teachers are essential for a group of learners to successfully learn content as well as language simultaneously. Being able to integrate these two aspects in a lesson is what is called 'CLIL skills'. It requires a number of different skills to do this effectively. The whole concept of good CLIL practice was divided in 6 categories (Dale 2010) or: "Keystone competencies", as we call them in this research.

The six key competencies of CLIL

The six keystone competencies of CLIL are the overlying entities under which we validate a number of criteria that are representative for each keystone competency. These keystone competencies are based on numerous studies generally concerning different aspects of "effective learning". (e.g. Mehisto, 2008; Dale, 2010; Reppen, 2002).

The six keystone competencies are described as methods for:

1. **Activation:** the general need for activation (Mehisto, 2008; Dale, 2010).
   This competency focuses on *activating prior knowledge* or *activating existing knowledge* when starting a lesson. In order to do this there are several techniques that a teacher can use, such as motivation which can be done by using visuals and activating language.

2. **Providing lesson input:** giving correct information which suits the age and level of learners and the time of the year. (Cummings, 2000; Krashen, 1983).
   Lesson input can be defined as '*the information used to help learners understand ideas and construct meaning'.* Learners learn to listen to, watch, look at or read input and use this input to carry out tasks or activities.

3. **Guiding understanding**: for a teacher to provide the right framework (or scaffold) for learning. (Wood, Bruner and Ross, 1976; Vygotsky,1978).
   This basically is a follow up on 'providing lesson input' and it deals with guiding understanding in the target language. Processing input helps learners to understand it better.

4. **Encouraging output**: how to effectively encourage learners to speak and write in the second language (Ur, 1996; Barnes, 1992; Mercer, 2000, Reppen 2002; Scrivener 2005; Mehisto, 2008). This role is about encouraging output, about getting learners to speak and write. Output can be defined as '*the production of language and content in the target language'*.

5. **Assessing learning:** Setting up activities that help learners to achieve the learning aims (Biggs, 2003). To understand the differences between assessment of learning and assessment for learning (TAR, 2002). Understand the need for different styles of assessment (Baker, 2006). Assessment and feedback are essential to all learning and this is why CLIL focuses on providing tools to do this efficiently. Feedback can be written or spoken and should be provided on a regular basis. CLIL makes a distinction between the assessment *of* learning and the assessment *for* learning.

6. **Using projects in CLIL:** the need for a balanced language and subject focus (Dale, 2010). CLIL offers a number of options for projects like cross-curricular and integrated projects, grouping learners for CLIL projects and WebQuests for CLIL. The method also provides practical ideas and characteristics for CLIL projects.

These short descriptions of the six key stone competencies in CLIL illustrate that it takes time and effort for a teacher to become competent in teaching CLIL. Books like “CLIL skills” (Dale, 2010) provide teachers with hands-on tips and advice on how to achieve this. However, it is difficult to improve your teaching practice when you do not receive any constructive feedback from peers or colleagues. Furthermore, providing feedback on someone’s teaching practice without a comprehensive reflection tool is equally hard. And thus, for now it often remains unclear whether teachers are able to implement the six competencies in their lessons and at which level.

Now that we have identified the 6 key-stone competencies of CLIL practice it is time to have a look at which form this feedback tool should preferably take.

A rubric styled feedback tool

In order to provide feedback on teacher CLIL competency, we suggest the use of a rubric styled observation instrument. The use of a rubric as a manner of providing feedback has several distinct advantages as described by Levi (2005) and Quinlan (2012). Quinlan makes the comparison between using a rubric and a ´performance list´. According to her research, rubrics have the following advantages:

- Rubrics provide an absolute standard or a benchmark
- Rubrics provide expectations about which aspects will be provided feedback on
- Rubrics provide information on the standards that need to be met
- Rubrics provide indications of where the users are in relation to goals
- Rubrics increase consistency in teacher ratings of performance, products or understanding

A simple ´performance list´ or old fashioned note keeping would not have sufficed as a tool for feedback on teacher CLIL competency as it does not provide constructive, reliable and reproducible feedback. The Rubric format however does offer these characteristics. Levi (2005) states that a rubric; *´provides timely, meaningful feedback´*, *´encourages critical thinking´*, *´facilitates communication with others´* and, *´helps us to refine our teaching skills´*. These are all qualities a constructive feedback tool should have and thus supports the choice for the rubric styled feedback tool.

Using a rubric is not only convenient it is also a thoroughly tested and effective format for teacher assessment. This is shown by Bodzin (2003) who found that using a rubric produced very reliable results in the pilot validation of their Science Teacher Inquiry Rubric (STIR). This implies that, provided the reflection tool is thoroughly validated, the rubric format at least can be used, and is being used, as a feedback instrument for teachers.

There are two main types of rubrics; the analytic rubric and the holistic rubric. The holistic rubric provides only limited feedback and is mainly used to get an overall ‘sense’ for what the subject is trying to accomplish (Mertler, 2001). In this research we use the analytic rubric as it provides the best way of assessing when a fairly focused type of response is required. (Nitkoe, 2001). We aim to use 5 criteria in our rubric to be as complete as possible, giving as much feedback as we can without the rubric becoming too cumbersome and difficult to use (Reddy, 2009). The final number of criteria used in any given key-competency depend on the number of criteria that are validated.

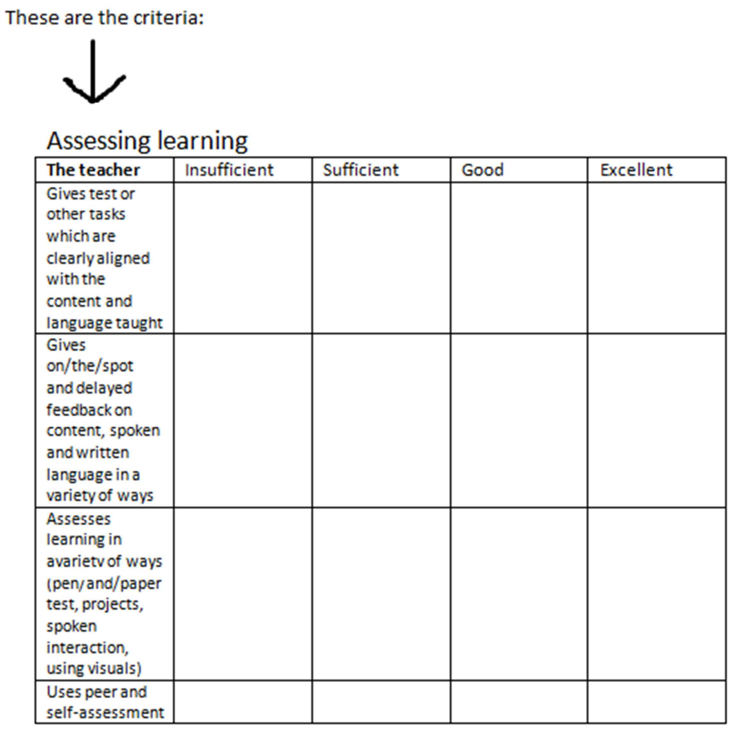
Parts of a rubric: Criteria

An important part of a rubric are its criteria (Fig 1). The criteria provide the backbone of the rubric and indicate which aspects are most important in that particular keystone competency. In our research the distinction of the six key-stone competencies is largely based on the work of Dale (2010). Consequently, our literature criteria correspond largely with the main criteria as described by Dale for each CLIL competency. However, one can't simply take criteria from the literature, use them in a rubric and expect the rubric to provide useful feedback for the teachers using it.

In order for a rubric to produce reliable results and give a genuine feel for the competency in CLIL skills the structure and descriptions of categories need to be validated (Allen, 2009). Furthermore, McNamara (1996) suggests in his research on second language performance that examining available literature to inform the content of a rubric is a good way of evaluating a rubric. This study also suggests collecting information from the academic literature and CLIL experts in order to further develop, determine, and test the rubric's specifications and scoring procedures.

The method that we used to validate the literature criteria of our rubric is largely based on the research of Levi (2005) and Baartman (2008). They suggest that collaborating to develop a rubric offers *"an opportunity to discuss shared goals and teaching methodologies''*, and lead to effective ways of evaluating and validating teaching practices.

In her research "*assessment of the wheel of competence*", Baartman makes use of an expert group to validate a mainly literature based rubric and finds this a very effective tool. This research aims to characterize *the design principles for a rubric styled feedback tool for good CLIL practice*. Therefore, a method as described by Baartman (2008) that incorporates the collaborative aspects as suggested by Levi (2005) seems to have a clear advantage over further expert/literature based validation (e.g. The Delphi method: North & Pike, 1969; Adler & Ziglio, 1996). Thus, in analogy of Baartman (2008) and Levi (2005) this research used an expert group of TTO teachers to validate the literature criteria for a rubric styled feedback tool.



These are the criteria:

↓

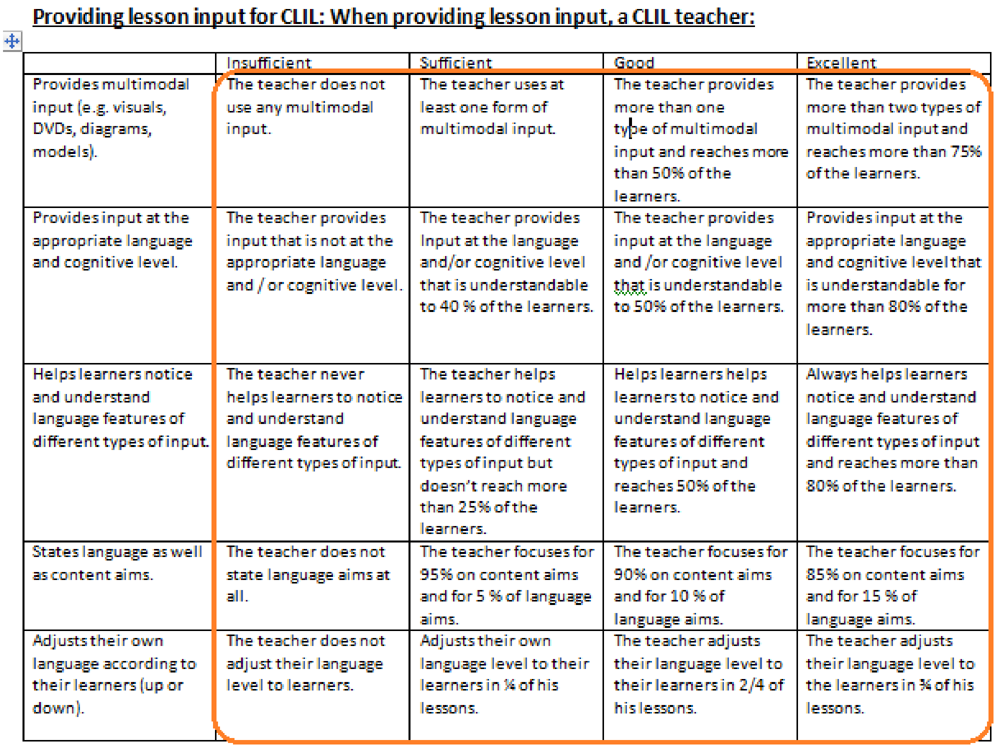
Assessing learning

The teacher	Insufficient	Sufficient	Good	Excellent
Gives test or other tasks which are clearly aligned with the content and language taught				
Gives on/the/spot and delayed feedback on content, spoken and written language in a variety of ways				
Assesses learning in avarietv of ways (pen/and/paper test, projects, spoken interaction, using visuals)				
Uses peer and self-assessment				

Fig 1: As shown in this <u>example</u> rubric, the Criteria are on the left hand side from top to bottom.

Parts of a rubric: Descriptors

A rubric is not complete without so called *descriptors (Fig 2).* The descriptors could be best described as the part of the rubric that specifies the characteristics of different criteria at different levels of competence. Like with the criteria, good and workable descriptors are a vital component of an effective rubric. Thus, determining what makes a good descriptor is very important. In order to determine design principles for these descriptors we asked our expert group to provide feedback on a list of descriptors that we designed. It is important to underline that unlike the criteria, which were distilled from academic literature, the descriptors are not directly based on any academic literature. However, we did take in to account the input provided by CLIL expert J. Skeet (2012, Personal communications) when designing these descriptors. The descriptors were created in such a way as to minimize value judgments, were clearly distinguishable in terms of achievement and left little room for interpretation in terms of desired output (Skeet 2012). The designed descriptors were used as a means to extract feedback from the expert group. Once more, following the approach of: "creating *an opportunity to discuss shared goals and teaching methodologies" (Levi 2005; Baartman 2008).* We asked our expert group to comment on the different descriptors provided, to give so called tips and tops and come up with concrete points for improvement. Transcripts of the different discussions and work sessions have been analysed and key design principles distilled from these transcripts.



Providing lesson input for CLIL: When providing lesson input, a CLIL teacher:

	Insufficient	Sufficient	Good	Excellent
Provides multimodal input (e.g. visuals, DVDs, diagrams, models).	The teacher does not use any multimodal input.	The teacher uses at least one form of multimodal input.	The teacher provides more than one type of multimodal input and reaches more than 50% of the learners.	The teacher provides more than two types of multimodal input and reaches more than 75% of the learners.
Provides input at the appropriate language and cognitive level.	The teacher provides input that is not at the appropriate language and / or cognitive level.	The teacher provides Input at the language and/or cognitive level that is understandable to 40 % of the learners.	The teacher provides input at the language and /or cognitive level that is understandable to 50% of the learners.	Provides input at the appropriate language and cognitive level that is understandable for more than 80% of the learners.
Helps learners notice and understand language features of different types of input.	The teacher never helps learners to notice and understand language features of different types of input.	The teacher helps learners to notice and understand language features of different types of input but doesn't reach more than 25% of the learners.	Helps learners helps learners to notice and understand language features of different types of input and reaches 50% of the learners.	Always helps learners notice and understand language features of different types of input and reaches more than 80% of the learners.
States language as well as content aims.	The teacher does not state language aims at all.	The teacher focuses for 95% on content aims and for 5 % of language aims.	The teacher focuses for 90% on content aims and for 10 % of language aims.	The teacher focuses for 85% on content aims and for 15 % of language aims.
Adjusts their own language according to their learners (up or down).	The teacher does not adjust their language level to learners.	Adjusts their own language level to their learners in ¼ of his lessons.	The teacher adjusts their language level to their learners in 2/4 of his lessons.	The teacher adjusts their language level to the learners in ¾ of his lessons.

Fig 2: An example rubric in which the descriptors are shown within the orange square.

Validation

In this research the term validation is mentioned frequently when dealing with the process of determining which literature criteria correspond with the views of our expert group on good CLIL practice. Validation thus refers to the process of finding a broader sense of agreement for certain statements (in our case the literature criteria). This process was conducted using a so called goodness of match test.

Research question

The concerns of our research school are explained and put in broader context in the theoretical framework. The main worry of the research school is the lack of an effective feedback tool for CLIL competencies. Because of the increasing amount of TTO Schools and teachers practicing CLIL skills, the lack of such a tool has become more apparent. The absence of a feedback tool for CLIL can be problematic for the quality of TTO education in the future. This research aims at the development of design principles for such a tool. To make a first step in the creation of a rubric styled feedback tool we validate literature criteria and identify key design principles for descriptors.

This problem is formulated in the main research question below. The sub-questions further narrow down our research question by specifically focusing on validating and/or determining the design principles for the two components of a rubric, namely: The *criteria* and *descriptors.*

- *What are effective design principles for a rubric styled instrument to assess CLIL competencies for TTO teachers?*

Sub Questions

1. *Which literature criteria are validated by our expert group, serving as means for providing constructive feedback on the CLIL practice of teachers?*
2. *What are the most important design principles for descriptors in a rubric for the production of feedback on teacher CLIL practice?*

Hypotheses

In a publication of her personal experiences, Mackenzie (2008) expresses her efforts to implement all aspects of CLIL. However, she underlines that even advanced teachers tend to stray occasionally from the practice as described by CLIL theory. Various reasons are mentioned: e.g. reluctance to change their approach/unwillingness to teach grammar. This, combined with findings from the initial meeting/discussion session with the expert group of TTO teachers in which they shared some ideas on their views on the application of CLIL theory, allowed us to come to a more focussed hypotheses for our first sub-question.

> **Sub question 1**) We expect to validate the majority of the literature criteria with possible lower degrees of validation in the competency: "*Encouraging speaking and writing*".

We had a discussion session with one of the CLIL experts at the Utrecht University (Skeet, personal communications, 2012). During this discussion the need for clarity of content and language use and non-judgmental writing when designing descriptors was made clear. Based on the findings of this discussion we hypothesise the following for our second sub-question.

> **Sub question 2**) We expect the expert group of TTO teachers to identify design principles for descriptors with a clear focus on language use and means for non-judgemental quantification of competencies.

Variables

In this research two very distinct variables were defined in relation to the two respective sub-questions. These variables are as follows:

Criteria

Criteria provide the backbone of a rubric and define the aspects that are deemed most important in any particular key-stone competency (Fig 1). The criteria used in this research were taken from the academic literature and validated by an expert group via a goodness of match analysis. This provided insight in the views of TTO teacher on good CLIL practice and the correspondence of these views with the literature.

Descriptors

Descriptors are the parts of a rubric that specify aspects of the different levels of competence for the various criteria (Fig 2). Aspects for effective descriptors were analysed through discussion sessions. This provided insight in the views of TTO teachers on essential attributes of good descriptors for a feedback tool for CLIL practice.

Personal relevance

We believe this research to be a valuable learning experience for anyone who wants to become a (TTO) teacher. By doing this research, we have gained a better understanding of the different perceptions one can have on effective CLIL skills. To be aware of pitfalls and opportunities in effective CLIL education is something every TTO/international teacher should work on and by having done this research we consider ourselves to have taken a very good first step.

Practical relevance

Numerous studies have been conducted trying to identify which aspects of feedback make it more or less valuable to the learning process (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Butler &Winne, 1995; Hattie &Timperley, 2007; Kluger&DeNisi, 1996). However, as was already shown by Page in 1958 (Page, 1958) they all seem to agree that receiving and giving feedback in one way or another is an essential part of this learning process. This research aims to make the first step in identifying design principles for a feedback tool for effective CLIL practice. Our research thus adds to the current body of knowledge on CLIL by making a first step in the creation of an effective feedback tool for CLIL practice. Consequently it provides a basis to help TTO teachers come to grips with the CLIL methodology more effectively.

We believe, that with the increasing pressure from the European platform on effective implementation of CLIL skills in the classroom, the findings presented in this research are very much a contribution to the modern TTO teaching community.

Methods & Materials

In this research a combination of quantitative and qualitative techniques are used to answer the main research question and sub-questions. In every aspect of our study the expert group of TTO teachers played an important role.

Respondents (The expert group of TTO teachers)

This research aims to gain insight into how different teachers define design principles for a rubric styled feedback tool for CLIL. Therefore, teachers from different subjects were included in our expert group of TTO teachers. By using teachers from different subjects, a too limited approach whilst validating the literature criteria was avoided (Skeet, personal communications, 2012). The teachers taught the following subjects: “Chemistry, History and English”. We aimed to have a relatively large expert group of TTO teachers (six TTO teachers at least) in order to increase the validity of the results. However, the research school was only able to provide three TTO teachers that fit the selection criteria during the course of this research. For a teacher to be included in the expert group of TTO teachers, they had to have experience with CLIL. Furthermore, they had to have substantial experience as a teacher and/or an explicit eagerness to work with CLIL and have a clear idea on how to implements this effectively.

The science teacher was a very experienced teacher who has worked in the U.K. and India for many years. The history teacher was an experienced teacher as well and works for the University of Leiden as a part-time researcher for Ancient History besides his regular teaching job. The English teacher was still at the start of her teaching career, but familiar with CLIL literature. Furthermore, she was very interested in the use of CLIL in her and other subjects and seemed to have concrete ideas about how CLIL should be implemented. Based on their knowledge of CLIL theory, their extensive teaching experience, enthusiasm for CLIL practice or both, these teachers were expected to provide valuable contributions to this research.

Instruments

In this study quantitative instruments were used to validate the literature criteria for a rubric styled feedback tool for CLIL competency (sub-question 1). Qualitative instruments were used to determine design principles for the descriptors (sub-question 2).

Quantitative instruments: Towards answering sub-question 1

In order to validate the literature criteria for each keystone competency, separate tables of comparison were designed for each competency (appendix 3). The tables of comparison contained literature based criteria (literature criteria) and criteria created by our expert group of TTO teachers. The number and type of criteria in each table of comparison varied per keystone competency as it was completely dependent on the number of criteria provided by the members of our expert group of TTO teachers.

Using tables of comparison as a basis for the so called “goodness of match tests” was inspired by a study from Baartman (2008). The goodness of match test measured to what extent two criteria tried to convey the same meaning/concepts. Using the tables of comparison, the expert group of TTO teachers could give scores between any given literature criterion and a criterion made by the expert group of TTO teachers. Scores were given in numbers between 1 and 10, with 1 meaning a lack of overlap in meaning or concept, and 10 meaning that the two criteria convey exactly the same meaning. In analogy with Baartman (2008) goodness of match scores of 6 and higher were interpreted as a validation of those specific literature criteria.

The validity of using the goodness of match test for the validation of assessment/feedback criteria is strongly supported by the uses of this method by Baartman (2008). In her thesis: "*Assessing the assessment*", she used the goodness of match tests successfully to test quality criteria for competence assessment programmes. However, contrary to the study done by Baartman (2008), an expert group of TTO teachers instead of CLIL experts was used in this research. Furthermore, this research used a relatively small expert group of TTO teachers making further statistical analysis impossible. These aspects of this research need to be kept in mind when interpreting the data.

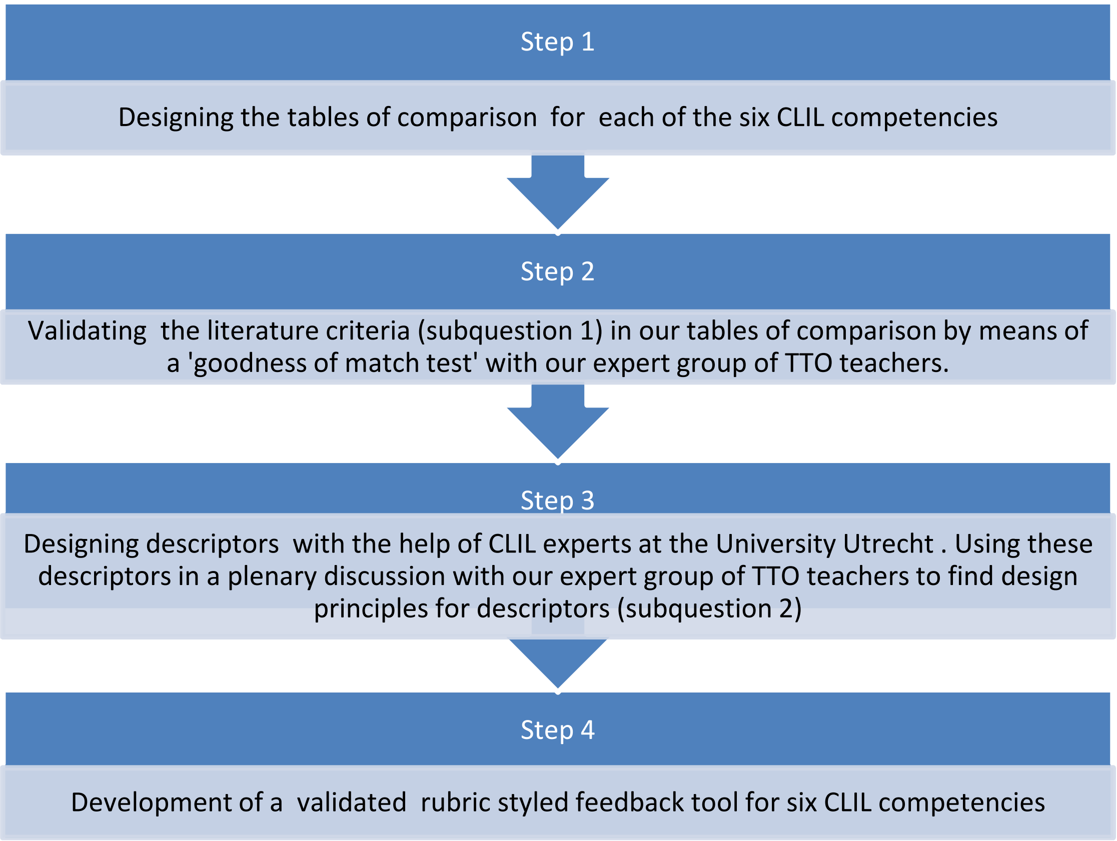
Qualitative Instrument: Towards answering sub question 2

This research aims to identify design principles for workable descriptors in a feedback tool for effective CLIL practice using an expert group of TTO teachers. As a tool to achieve this, we designed descriptors for the expert group of TTO teachers to provide feedback on. The descriptors were discussed with and approved for this purpose by Skeet (CLIL expert at Utrecht University). The feedback provided by our expert group of TTO teachers was recorded on tape and notes were made during the feedback session. In analogy with Bogdan (1982) the videotapes were used to support the notes from the discussion. We used the '*analysing writing*' approach (Coffey 1996) and subsequent '*grounded coding*' (Taylor 2010) to structure the findings.

When using grounded coding, the first step is for the researchers to set aside his/her existing knowledge on aspects for good descriptors and looks for "new" themes produced by the expert group of TTO teachers in the data. We did this by checking each other throughout the coding process and helping each other to keep an open mind towards the data. We combined the data for themes, ideas and categories of responses. Furthermore, similar passages of feedback were marked with a code label so that they could easily be retrieved at a later stage for further comparison and analysis. Using this method (Taylor, 2010) made it easier to search the data, make comparisons and to identify patterns.

The process of grounded coding has been an effective way to order an amount of qualitative data, into an organized unbiased overview of categories. By implementing the method described by Taylor (2010) we organized the qualitative data into three categories: *language, measurability and attainability of the descriptors*.

Setup of the research



Step 1

Designing the tables of comparison for each of the six CLIL competencies

Step 2

Validating the literature criteria (subquestion 1) in our tables of comparison by means of a 'goodness of match test' with our expert group of TTO teachers.

Step 3

Designing descriptors with the help of CLIL experts at the University Utrecht . Using these descriptors in a plenary discussion with our expert group of TTO teachers to find design principles for descriptors (subquestion 2)

Step 4

Development of a validated rubric styled feedback tool for six CLIL competencies

The collection of data for this research was carried out as follows:

Step 1: The main literature criteria for each CLIL competency were identified using CLIL skills (Dale 2010) and meetings with J. Skeet (CLIL experts at the Utrecht University). The expert group of TTO teachers created their list of criteria for each CLIL competency. With the combined information we made the tables of comparison (appendix 3).

Step 2: The expert group of TTO teachers carried out the ´goodness of match test´ between the literature criteria and the criteria made by the expert group of TTO teachers (appendix 1).

Step 3: In a plenary discussion we collected feedback from the expert group of TTO teachers on the essential characteristics of good descriptors. The findings were used to produce three concrete design principles for descriptors (results).

Step 4: The collected data was used to create a validated (criteria only) rubric styled feedback tool for all the six CLIL competencies (appendix 4; sample).

Answering the sub questions

In this research answering the two different sub-questions required two different approaches. We met with the expert group of TTO teachers on two different occasions. During the first meeting the expert group of TTO teachers provided criteria for the six keystone competencies and used these in a goodness of match test to validate the literature criteria (sub question 1). During the second meeting the expert group of TTO teachers participated in taped plenary discussions about the quality of the design principles for workable descriptors in a feedback tool for CLIL. These discussions were based on the previously designed descriptors.

Validation of literature criteria (sub-question 1)

In analogy with the method as described by Baartman (2008), the expert group of TTO teachers wrote down the criteria they thought were important for each keystone competency of CLIL. To prevent influencing the expert group of TTO teachers, only a general introduction about the feedback instrument and its criteria was provided.

The expert group of TTO teachers were given ten minutes to enter as many criteria as they could in an empty rubric. They were then asked to review the criteria entered and if possible add two more criteria not yet included. This resulted in six lists of criteria, one for each keystone competency. The lists were reviewed in order to combine duplicate and comparable criteria. The resulting lists were discussed in a subsequent plenary session in order to achieve mutual understanding of the criteria and to generate an even more workable list of criteria. This process solved the problem that different words are often used for the same idea.

We analysed the results from our goodness of match test between the literature criteria and the criteria designed by the expert group of TTO teachers. For this analysis we used the mean value for the goodness of match test from all three members of the expert group for each criterion. The process of validating the literature criteria is depicted in the flowchart presented below.

The process of validating literature criteria

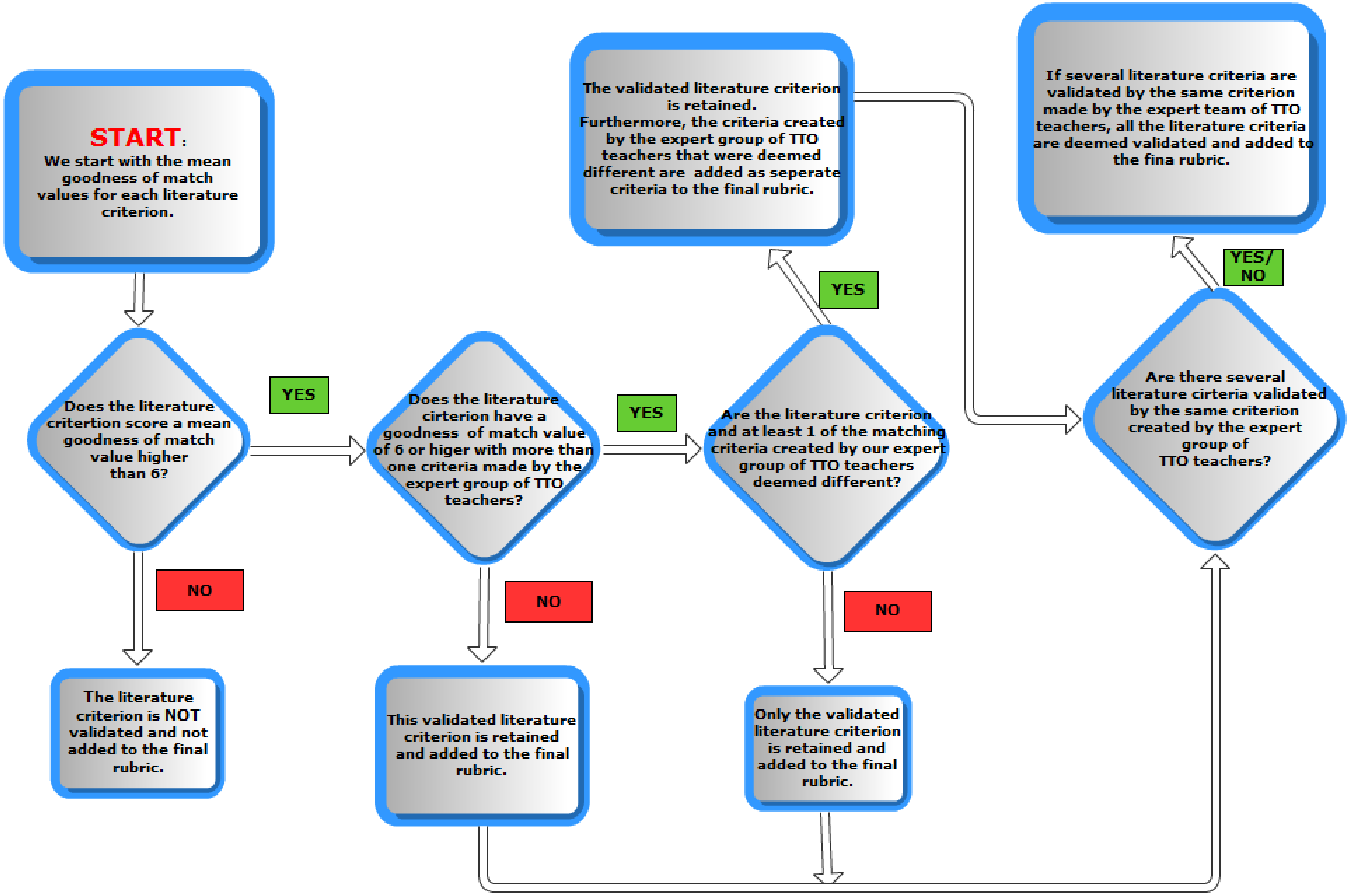

START: We start with the mean goodness of match values for each literature criterion.
Does the literature critertion score a mean goodness of match value higher than 6?
YES
NO
The literature criterion is NOT validated and not added to the final rubric.
Does the literature cirterion have a goodness of match value of 6 or higer with more than one criteria made by the expert group of TTO teachers?
YES
NO
This validated literature criterion is retained and added to the final rubric.
Are the literature criterion and at least 1 of the matching criteria created by our expert group of TTO teachers deemed different?
YES
NO
Only the validated literature criterion is retained and added to the final rubric.
The validated literature criterion is retained. Furthermore, the criteria created by the expert group of TTO teachers that were deemed different are added as seperate criteria to the final rubric.
Are there several literature cirteria validated by the same criterion created by the expert group of TTO teachers?
YES/ NO
If several literature criteria are validated by the same criterion made by the expert team of TTO teachers, all the literature criteria are deemed validated and added to the fina rubric.

Fig:3: A flowcharts representing the process of including certain literature criteria and/or criteria made by the expert group of TTO teachers in de final rubric styled assessment tool. One starts at: “Start”, and follows the appropriate arrows/answers through the scheme.

In the final representation of the findings from this research, the number of criteria that were validated are shown per CLIL competency. Furthermore, the average degree of validation (the closer to 10 the higher the degree of the overall validation) for all validated literature criteria in each of these CLIL competencies is shown.

The calculation of the average degree for goodness of match of all the validated criteria within a given keystone competency, deserves a little more explanation. Averages were calculated using only the validated criteria (matches that got a 6 or higher). When more than one criterion created by the expert group of TTO teachers validated the same literature criterion, only the strongest match was incorporated in the final calculation of the average goodness of match for that CLIL competency as a whole.

The reliability of the goodness of match analysis as used in this study for the validation of criteria for teacher assessment is supported by findings from Baartman (2008). However, we used a relatively small group of TTO teachers in this study compared to a relatively large group of CLIL experts in the study conducted by Baartman (2008). This marked difference must be kept in mind when interpreting the data.

Identifying important design principles for descriptors (Sub-question 2)

By means of two plenary discussions with our expert team of TTO teachers we identified design principles for descriptors in an effective CLIL feedback tool. These sessions were started by introducing the general purpose and use of rubrics as a means for providing feedback. The main aspects of working with rubrics were explained. Furthermore, it was emphasized that the expert level was not expected to be attained for every, if any, criteria. We explained what criteria and descriptors are and what their purpose was in the rubric. Moreover, it was underlined that not all the rubrics for each of the six CLIL competencies had to be used in one single lesson.

In the second meeting with the expert group of TTO teachers the quality of the descriptors of the six rubrics was discussed. The expert group of TTO teachers provided so called tips and tops (positive and negative remarks) for every set of descriptors of each criterion. We facilitated the discussion and made notes of their key findings during the discussion. The sessions were recorded on video and transcribed afterwards. The transcribed data was used to check and complement the notes made during the discussion.

We qualitatively analysed the notes and transcript of the video recording we made during the plenary discussion session. During this qualitative analysis the main arguments were coded. This made it easier to search the data for specific comments, make comparisons, and identify patterns in the data. For this, we used the previously described grounded coding (Taylor, 2010). The coding was done by all members of the research team individually before comparing and combining the findings. This resulted in three distinct categories of feedback with regards to design principles for workable descriptors in a feedback tool for CLIL, namely: *The language used when describing the descriptors, the measurability and the attainability of the descriptors*.

Results

Validation of literature criteria (sub-question 1)

In this section we present the findings with regards to the goodness of match test that was used to validate the literature based criteria.

A graphical representation of the results can be found in the graph shown below. The structure of this graph is further explained in the subscript accompanying it (Fig 4).

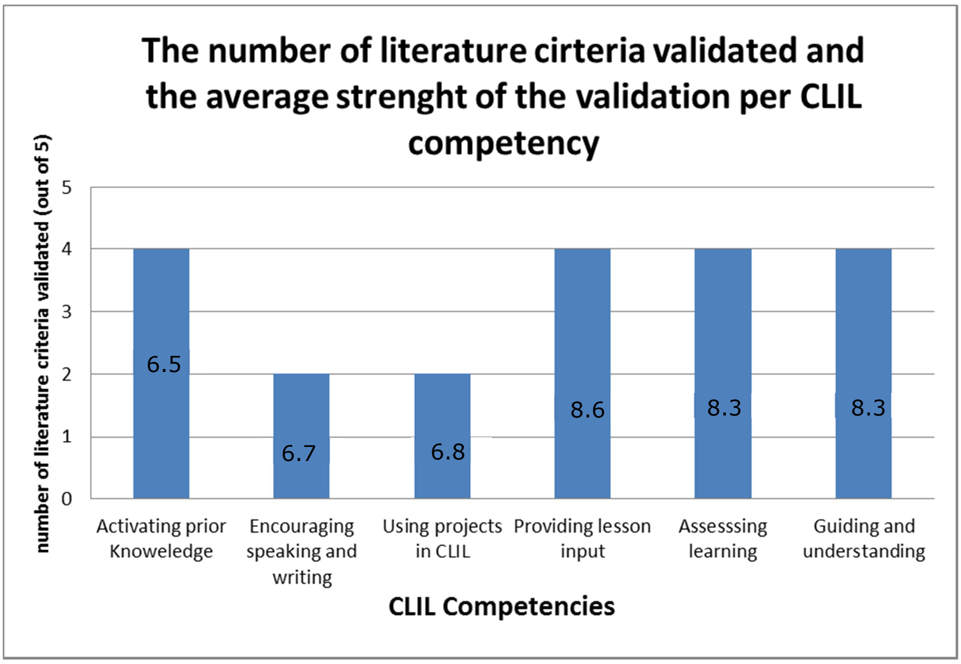


Fig 4: This graph shows the number of criteria (out of 5) that have been validated by our expert team based on the goodness of match results. Furthermore, shown inside the bars is the average degree of the goodness of match test for al validated criteria within that CLIL competency. When multiple expert criteria validated the same literature criteria only the strongest match was incorporated in the calculation.(see methods).

This research shows that for the CLIL competencies: *Activating prior knowledge*, *providing lesson input*, *assessing learning* and *guiding and understanding*, 4 out of 5 literature criteria have been validated through a goodness of match test of 6 or higher.
For the CLIL competencies: *Encouraging speaking* and *writing and using projects in CLIL*, only 2 out of 5 literature criteria were validated (Fig 4).
The highest degree of validation was found in the CLIL competencies: *Providing lesson input*, *assessing learning* and *guiding and understanding*. All these competencies scored an 8 or higher on average.
In the categories: *Activating prior knowledge*, *encouraging speaking and writing* and *using projects in CLIL*, the average degree for the goodness of match test was lower with a score between 6 and 7.
This research shows that the majority of the literature criteria were validated (20 out of 30). However, the expert group of TTO teacher never validated all the literature criteria in any of the keystone competencies of CLIL we took from the literature (Fig 4; Appendix1).
This research aimed to find out: W*hich literature criteria are validated by an expert group of TTO teachers, serving as means for providing constructive feedback on the CLIL practice of teachers?* The results indicated that this research has been successful in identifying these literature criteria.

Validity and reliability of results (sub- question 1)

The method used to validate the literature criteria has been proven to be reliable by Baartman (2008). However, this research used Baartman's research approach under different circumstances. Therefore, we felt that our findings had to be given an extra, separate validation to make them more reliable. This was done by asking feedback on the data of the expert group of TTO teachers. They concluded that the goodness of match test had successfully captured the overall intentions of the expert group of TTO teachers. Members of the expert group of TTO teachers expressed that the right criteria were included in the final rubric. This validation gives extra strength to the presented results. The reader has to keep in mind that the findings of this research only represent the input of the three people of which our expert group of TTO teachers was made up. This has most likely influenced the validity of the results.

Identifying important design principles for descriptors (Sub-question 2)

This section provides the most important findings of this research on *identifying important design principles for descriptors in a feedback tool for CLIL practice*.

The main feedback provided by our expert group of TTO teachers was taken from the video transcripts and the notes made by the researchers during the group discussions. The provided feedback was labelled following the methodology of "grounded coding" (methods). This process produced three categories of feedback. These categories of feedback and the general findings for each category are:

1. **Language:**

   TTO teacher expert 1: *''The way some of the descriptors are written leaves a lot of room for interpretation. This is not what I would like when working with a feedback tool that wants to provide precise feedback on teacher CLIL practice.''*

   The way the descriptors are written must be very specific. A number of descriptors contained terminology like: "*in a number of ways*" or "*often*". This was deemed confusing and unclear for teachers. Our expert group of TTO teachers also underlined the need for clear and precise wording and phrasing when designing descriptors as this would further decrease the risk of multiple interpretations. Clarity, with the risk of occasionally being judgmental to a degree, was preferred over vagueness.

2. **Measurability:**

   TTO teacher expert 2: *'' Some of the descriptors that you give in your rubric are too difficult to measure in a classroom situation. How can I see whether 60% or 80% of the class is really paying attention? This needs to be improved upon.''*

   A lot of descriptors progressed through the different levels of competence by using phrases like: "*increasing levels of student participation*" or "*Increasing levels of student comprehension*". Our expert group of TTO teachers pointed out that in practice such wording would make it hard if not impossible to measure the progress accurately and need a more workable alternative.

3. **The attainability:**

   TTO teacher expert 3:*''When I first looked at the rubrics and the standards set to become an expert teacher, I despaired a little. This is impossible! The standards used seem unreachable and need to be more realistic for this rubric to be useful to teachers.''*

> Our expert group of TTO teachers expressed the feeling that the standards of our rubrics might have been too high and thus off-putting. More specifically the expert group of TTO teachers felt that the level of an 'expert teacher' was almost always unattainable or even unrealistic. This should not be the case in a constructive feedback tool and needs improvement.

A further analysis of the feedback provided by the expert group of TTO teachers led to three very concrete design principles for descriptors of a feedback tool for CLIL.

- Descriptors should be written using words that make sure that there is little or no room for multiple interpretations.
- Descriptors should only contain measurable units and these units should be made very explicit. (E.g. "*percentage of learners that are paying attention*" should *not* be used in a descriptor).
- Descriptors should always describe realistically attainable levels of competence.

Two more general points of feedback were provided by the exert group of TTO teachers. Although these points were not identified as separate categories, this feedback did carry valuable information and was used to create the following general concepts to keep in mind.

- Descriptors should be designed to be applicable for the work- and textbooks used in that particular school.
- Language teachers can only work with rubrics and descriptors that fit their subject, as the focus between subject and language overlaps completely in their subject.

This research aims to find what the *most* important design principles for descriptors in a rubric for the production of feedback on teacher CLIL practice are. The results presented above indicate that this research has been successful in determining three clear design principles for descriptors.

Validity and reliability of results (sub- question 2)

As with the validation of our goodness of match results for the criteria, the findings of this analysis were also discussed with our expert group of TTO teachers. This process showed that the findings were representative of the ideas of the expert group of TTO teachers and thus, adds to the credibility of the results. Despite that our data analysis seems quite strong, it still only represents the feedback of three people. And just like with the goodness of match results this should be kept in mind, as it's likely to have influenced the validity of the presented findings.

Conclusions

Validation of literature criteria (sub question 1)

This study shows that the majority of the literature criteria were validated by the expert group of TTO teachers (20 out of 30). This finding supports the first part of the hypothesis for sub-question 1. The amount of literature criteria that are validated and the degree of validation vary for each of the six keystone competencies. The expert group of TTO teachers seemed to agree with the literature's idea of good CLIL practice for the competencies: "*Providing lesson input*, *assessing learning* and *guiding and understanding*". The expert group of TTO teachers had different opinions on good CLIL practice while validating criteria for the competencies:'' *Activating prior knowledge*, *encouraging speaking* and *writing and using projects in CLIL*.'' This supports the second part of the hypothesis for sub-question 1. In analogy with Mackenzie (2008), we only expected a decrease in validation of the competency: "*Encouraging speaking and writing*".

We believe that less literature criteria were validated in the above mentioned keystone competencies due to the fact that our expert group of TTO teachers produced criteria that tended to have a different emphasis than the literature criteria. The criteria provided by the expert team of TTO teachers were generally very focussed on 2 or 3 very specific aspects of a CLIL competency. The criteria found in the literature generally covered every aspect of the CLIL competencies. These findings lead us to conclude that despite the extensive (intuitive) knowledge with regards to practicing effective CLIL skills, there are parts of the CLIL literature used in this study that might need a better introduction to be recognized as valuable/desirable by our expert group of TTO teachers. The inability of our expert group of TTO teachers to provide a number of varied criteria in specific keystone competencies, suggests that there are areas of CLIL where there's still a lot to gain for TTO teachers. This is a valuable insight as it might help TTO teachers to work on their ability to practice effective CLIL skills.

Identifying important design principles for descriptors (Sub question 2)

Based on the results from this study we conclude that three main categories of design principles for descriptors can be defined, namely: ***Language, measurability and attainability***. This in part supports the hypothesis for sub-question 2, as a clear focus on language use was present. Furthermore, it is in accordance with findings by H.G. Andrade (Andrade, 1997; Andrade, 2000) who also pointed out that clear language and measurability are important aspects of good descriptors.

In terms of the categories: "*measurability and attainability*", it was interesting to note that our expert group of TTO teachers expressed a need for a different style of quantification of descriptors than used by the researchers. The expert group of TTO teachers stated that if it contributed to the overall clarity, descriptors could be judgmental to an extent. This was not in accordance with previous findings by Andrade (1997, 2000). Furthermore, we did not expect the category of: "attainability" as a design principle for descriptors. These unexpected findings are considered very valuable as they help us understand teacher's perspective with regards to what is essential for the development of workable descriptors.

This research has made a good first start in identifying effective design principles for a rubric styled instrument to assess CLIL competencies for teachers. In accordance with previous findings by Andrade (Andrade, 1997; Andrade, 2002) we identified ''*language*" and ''*measurability''* as essential aspects of good descriptors. Furthermore, in comparison with the findings from Andrade, we identified the new category ''*attainability''*.

Being aware of which aspects are important when designing descriptors for a rubric can help teachers in optimizing their own rubrics for good CLIL practice.

Based on the above, we consider the main research question answered to the extent this could be expected from this preliminary research.

Discussion

As shown by our results, this research has provided a first stepping stone for developing a tool that supports teachers to improve their CLIL practice. The need for creating such a tool cannot be underlined enough. Several studies show that without constructive feedback, learning is generally not optimal (Page, 1958; Hattie &Timperley, 2007). In the study performed by Baartman (2008) the expert group was relatively large and the amount of data collected sufficient to buffer for statistic anomalies.

However, this research was performed under less than ideal circumstances. Our expert group of TTO teachers only consisted of three teachers. Furthermore, the research was not carried out in the planned three sessions of 3-4 hours but had to be done in three sessions of only 1-2 hours. This greatly influenced the amount and the quality of the data collected and made it impossible to make the data more reliable by means of statistical analysis.

The expert group of TTO teachers often didn’t provide as many and more importantly as varied criteria and design principles for descriptors as expected. The criteria and comments on design principles provided by our experts often came into focus at the beginning of a session and then often came back to the same points/criteria throughout that session. This dynamic further limited the amount of data collected.

Due to the limiting factors of this research, personal views, levels of experience with CLIL, levels of understanding of the assignment and varying levels of teaching experience are likely to have had substantial effect on our findings. This notion was supported by the core data with regards to the validation of the literature criteria. Here it was clear that levels for goodness of match varied extremely (e.g. 10, 1, 1) between members of the expert group of TTO teachers. For the descriptors the same holds true as very opinionated members of our expert group of TTO teachers could easily influence final arguments and were more likely to influence their fellow members due to the small size of the group.

All the above supports the idea that the presented degree of validation needs to be viewed critically. Based on the findings, it is not necessarily true that the literature criteria that are validated are in fact good criteria and vice versa. The same goes for the descriptors, as it’s unlikely that all the essential design principles have been identified by this research. The opinions and feedback presented by the expert group of TTO teachers might have been nuanced if we would have used a larger expert group of TTO teachers.

Making sure the final product is thoroughly validated is important. This is indicated by Roblyer and Wiencke (2003) who show that only a thoroughly validated rubric can provide reliable feedback and a good means to encourage interaction by students and instructors.

We believe that despite the obvious flaws, this research has made a good first attempt at indicating possible valuable and less valuable literature criteria for all the six CLIL keystone competencies. This notion is supported by the fact that teachers at the research school have expressed the desire to work with and further develop the product. They recognized the validated criteria and design principles for descriptors from this study as valuable steppingstones to a practical way of improving their daily teaching practice.

Recommendations

Practical Recommendations

The members of our expert group of TTO teachers were very enthusiastic about the idea of having a feedback tool for CLIL. They confirmed our general conclusion to sub-question 1: "*that there's still a lot to gain for teachers in terms of their ability to practice effective CLIL skills.* " They suggested that in order to do this effectively in practice, the instrument should strictly be used in an open relation between the observer and the observed. It must be clear in which class the feedback tool will be used, as it should not become a "surprise assessment". This fits well with an additional point made, namely the observer and the observed should agree in advance on which part of the CLIL spectrum to focus as it was deemed unrealistic to analyse and put in to practice all aspects of CLIL in one lesson.

With regards to our findings in answering sub-question 2: it was suggested that although categories were defined effectively, the rubrics need to be fully revised using the three suggested design principles for the descriptors. Furthermore, it was suggested that since the observed teacher is expected to be absorbed by teaching, the observer should always keep a (written) record of why certain descriptors are contributed to the teaching practice observed. Simply receiving the rubric back without argumentation could lead to disagreement and take away from the learning process.

Suggestions for further research

Due to the setup of this research, especially considering the size of the expert group of TTO teachers, we believe the results presented should be regarded as preliminary. Further research should thus focus on an increase in size of the expert group of TTO teachers to produce results that can be tested statistically adding validity to the final product. This research relied largely on the concept of CLIL as defined by Dale(2010). However, there have been other experts who used different ways of structuring CLIL competencies and therefore might have formulated slightly different "literature criteria" (e.g. Mehisto 2010). The use of such criteria should be considered in subsequent studies as it is expected to add depth to the final product.

This study only aimed to identify the categories of design principles for descriptors of CLIL. Using a larger expert group of TTO teachers, this research could be repeated with an explicit focus on producing the actual descriptors for all the validated criteria. This could result in a completely validated instrument.

The results presented in this research suggest that the familiarity of TTO teachers with different parts of the theoretical CLIL spectrum varies greatly. Based on these findings we suggest a study that aims to describe more specifically for which CLIL competencies the theory is best known to TTO teachers. Such a study could also try to find out why specific parts of the CLIL theory are better known to TTO teachers. Findings from such a study could make it clear where an additional focus on CLIL theory is needed and how to do this effectively. This knowledge could improve efficiency when educating TTO teachers.

Reflections

We experienced doing this research as a very challenging and insightful experience. Regarding the process, we now realize that working with teachers brings different challenges than working with students. Since the research was conducted towards the end of the school year, the teachers on our research school were often very occupied and reluctant to devote too much time to things outside their already busy working day. This led to a substantial reduction in the desired expert group size of TTO teachers and the amount and length of the sessions we were able to arrange with them. If we had anticipated this beforehand this could have improved the validity of our results.
This research taught all members of the research team to work with new ways of looking at data. The combination of quantitative and qualitative methods was initially met with reservations, but was found to provide the flexibility needed to come to insightful conclusions and practical recommendations.

In terms of the product, doing this research underlined the differences in backgrounds between the members of our research team and it would not be an exaggeration to say it was a struggle at times. However, it was in this struggle that a lot of learning took place and mutual understanding of individual views did increase through discussions. The process of listening to each other and the members of our expert team of TTO teachers with the intention to truly understand the messages they were trying to convey was essential to the success of this research. It is a skill that we consider of great use and value in our future careers.

Literature list

Adler, M. & Ziglio. (1996). *Gazing into the Oracle: the Delphi Method and its Application to Social Policy and Public Health*. London: Jessica Kingsley Publishers.

Allen, E. & J.Seaman, (2009). Learning on demand. Online article: Babson Survey Research Group.

Andrade, H.G. (1997). Understanding rubrics. *Educational leadership*, 54 (4), 14-17

Andrade, H.G. (2000). Using rubrics to promote thinking and learning. *Educational leadership.* 57 (5), 13-18

Baartman, L.K.J., (2008). *Assessing the assessment. Development and use of quality criteria for Competence Assessment Programmes.* Utrecht: Print Partners Ipskamp.

Baker, W. (2006). *Studies in Second Language Acquisition.* Cambridge: Cambridge University Press.

Bangert-Drowns, Kulik, Kulik, & Morgan. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research* (2) 213-238.

Barnes, R. (1992). *Successful study for degrees*. London: Routledge.

Biggs, J. (2003). *Teaching for Quality Learning at University.* Berkshire: Open University Press.

Bodzin, A.M. & K.M. Beerer. (2003). Promoting inquiry-based science instruction: The validation of the Science Teacher Inquiry Rubric. *Journal of Elementary Science Education* (2) 39-49.

Bogdan, R.C. (1982). *Qualitative research for education: an introduction to theory and methods.* London: Pearson.

Brookhart, S.M. (2008). *How to give effective feedback to your students?* Alexandria: Association for supervision and curricula development.

Butler, D.L. & P.H. Winne. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Education Research* (3) 245-281.

Coffey, A. & P. Atkinson (1996). *Making sense of Qualitative data.* Thousand Oaks: Sage Publications

Cummings, C. (2000). Winning Strategies for Classroom Management. Alexandria: Association for Supervision and Curriculum Development.

Dale, L., Van der Es, W., & Tanner, R. (2010). *CLIL Skills.* Leiden: GrafiMedia.

Gibbs. G.R. & C. Taylor. (2005). *How and what to code.* University of Huddersfield: Online QDA Web Site.

Hattie, J. & H. Timperley (2007) The power of feedback. Review of Educational Research (1) 87-122.

Kluger, A. & A. DeNisi*. (1996).* The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* (2) 254-284.

Krashen, S.D. (1983). *The natural Approach: language acquisition in the classroom*. San Francisco: The Alemany Press.

Levi, A.J. & Stevens, D.D. (2005). *Introduction to rubrics. An assessment tool, to save grading time, conveys effective feedback and promote student learning*. Sterling: Stylus Publishers.

Mackenzie, A. (2008). How should CLIL work in practice? *(One stop English). Number one for English language teachers*. Retrieved August 02, 2012, from: http://www.onestopenglish.com/support/methodology/teaching-approaches/how- should-clil-work-in-practice/156531.article

Marenzi, I., Kupetz, R., Nejdl, W., & Zerr, S. (2010). *Supporting active learning in CLIL through collaborative search. Informally published manuscript*, University of Hannover.

McNamara, T. (1996). *Measuring second language performance.* Oxford: Oxford University Press.

Mehisto, P. & M.J. Frigols. (2008). *Uncovering CLIL: Content and language integrated learning and multilingual education.* Oxford: Oxford University Press.

Mercer, N. (2000). *Scaffolding the development of effective collaboration and learning.* Cambridge: Cambridge University Press.

Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25)

Nitko, A. J. (2001). *Educational tests and measurement, an introduction*. New York: Harcourt Brace Jovanovich.

Quinlan, A.M. (2012). *A complete guide to rubrics.* Plymouth: Rowman & Littlefield Education.

Page, E. B. (1958). Teacher comments and student performance: A seventy-four classroom experiment in school motivation. *Journal of Educational Psychology*, 49(2), 173–81.

Reddy, Y.M. & H.G. Andrade. (2010). 'A review of rubic use in higher education'. *Assessment and Evaluation in Higher Education* (4), 435-448.

Reppen, R. (2002). *What does frequency has to do with grammar teaching?* Cambridge: Cambridge University Press.

Roblyer, M. D., & Wiencke, W. R. (2003) Design and Use of a Rubric to Assess and Encourage Interactive Qualities In Distance Courses. *The American Journal of Distance Education* 17(2): 77–97

Scrivener, J. (2005). *Learning Teaching: A Guidebook for English Language Teachers* . Ismaning: Hueber Verlag GmbH & Company.

Skeet, J. (2008) *Going Dutch! Acculturation of Pupils aged 15 to 19 at Dutch International Schools.* Utrecht: University of Utrecht.

Taylor, C and Gibbs, G R (2010) "How and what to code", Retrieved August 04, 2012, from: *Online QDA Web Site*: onlineqda.hud.ac.uk/Intro_QDA/how_what_to_code.php

Ur, P. (1996). *A course in language teaching. Practice and theory*. Cambridge: Cambridge University Press.

Vygotsky, L. &M. Gauvain, M. Cole. (1978). *Interaction between learning and development*. Cambridge: Cambridge University Press.

Wood, D., J.S. Bruner & G. Ross. (1979). The role of tutoring in problem solving. *Journal of child psychology* (2) 89 -100.

Appendix 1 (preliminary analysis: Validating literature criteria)

When we analyze the data in the manner stated in material and methods, the following interesting findings can be elicited from the data.

With regards to the competency:" **Activating Prior Knowledge**"

When we look at the analysis of the data as provided by our expert team we see that literature criterion **A** (*Helps learners to activate their existing subject knowledge and experience in a variety of ways.*) Has a marginal match (>6 but <7) with expert criteria **1** and **3** (*Introduces New Topic (content). Links to what is already known and provides context /connects to other topics" and "Engages students in various ways with focus on language/speaking: e.g. via Brainstorm " respectively.*

Literature criterion **B** (*Helps learners to activate their existing language knowledge and skills in a variety of ways*.) has relatively strong matches (>7) with expert criterion **3** and **4** (*Checks availability of necessary vocabulary*). Literature criterion **C** and **D** both only match (>6 but <7) with expert criterion **3**. Literature criterion **E** has no significant matches with criteria created by our expert group of TTO teachers it not validated by this research.

Interesting to note is that literature criterion **A,B,C** and **D** all had significant matches with expert criterion **3.** The above supports the inclusion of literature criterions **A,B,C** and **D** in the final rubric. Furthermore, expert criterion 1 was deemed redundant with parts of literature criterion **A** and **B** leaving only expert criterion **4** to be additionally included in the final rubric.

Activating prior Knowledge					
EXP/Lit-->	**A**	**B**	**C**	**D**	**E**
1	**6**	4	4,67	1,67	3,67
2	4,3	3	2	3,33	3
3	**6,0**	**7**	**6,67**	**6**	4,33
4	3,0	**7,33**	5	3,33	3,33

Presented in this graph, are the averages of the goodness of match analysis as conducted by our expert team for the competency: "Activating prior knowledge". The Expert (Down) and literature (right) criteria are, for clarity reasons, replaced by boldly printed letters and numbers respectively. Original criteria can be found in appendix (nr). Significant matches (6 or higher) are printed bold and red

With regards to the competency:" **Encouraging speaking and writing**"

When we look at the analysis of the data as provided by our expert team we see that the input from our expert team has not provided validation for literature criterion **A, B and D** (**A**: "*Provides learners with plenty of speaking and writing opportunities, ensuring that learners always use English"*. **B**: "*Sets up pair and group work where learners communicate in English"*. **D**: "*Scaffold output, e.g.: by providing learners with speaking or writing frames"*.) Literature criterion **C** has a marginal match (>6 and <7) with expert criterion **3** (*Teacher makes writing task (language/content) integral part of class*). Literature criterion **E** (*Stimulates and challenges learners to produce more complex output)*shows marginal matches (>6 and <7) with expert criterion **2** and **3** ("*Teacher sets up moments for student presentations*" and *Teacher makes writing task (language/content) integral part of class."* respectively*)* and a relatively strong match (>7) with expert criterion **5** (*Teacher uses activating work formats (authentic role-play*). The above supports the inclusion of literature criterion C and E in the final rubric and as Expert criterion 2 and 3 were both deemed to be well represented by literature criterion **E**, amongst others, this only left expert criterion **5** to be included separately in the final rubric.

	Encouraging speaking and writing				
EXP/Lit-->	**A**	**B**	**C**	**D**	**E**
1	5.3	2.7	2.7	4.3	4.3
2	4.5	5.0	2.0	3.7	**6.3**
3	5.0	1.0	**6.3**	5.0	**6.0**
4	5.0	1.0	3.7	2.7	4.0
5	3.5	4.5	3.7	4.3	**7.0**

Presented in this table, are the averages of the goodness of match analysis as conducted by our expert team for the competency: “Encouraging speaking and writing”. The Expert (Down) and literature (right) criteria are, for clarity reasons, replaced by boldly printed letters and numbers respectively. Original criteria can be found in appendix (nr). Significant matches (6 or higher) are printed bold and red.

With regards to the competency:” **Providing Lesson input**”

When we look at the analysis of the data as provided by our expert team we see that for literature criterion **A** (“*Provides multimodal input, e.g. visuals, dvds, diagrams, models”*) we found a perfect match with expert criterion **1** (=10). We find a strong match (>8) between both Literature criterion **B and E** (“*Provides input at the appropriate language and cognitive level” and Adjusts their own language according to their learners (up or down)* respectively*)* and expert criterion**3** (“*Language used by teacher is clear. Teacher Has wide vocabulary range and provides/uses necessary jargon”)***.** Literature criterion **C** shows a good match with expert criterions **2** (“*Focuses on helping students to understand input. Attention for used grammar and vocabulary”; >7*)and a marginal match with expert criterion **3** (>6 and <7). For literature criterion **D** (“*States language as well as content aims.”*) no validation was found in our goodness of match analysis. The above supports the inclusion of literature criteria **A**, **B**, **C** and **E** in the final rubric. The only multiple matches between literature and expert criteria were those between literature criteria **C** and expert criteria **2** and **3**. However, expert criteria **2** and **3** were deemd to be covered in literature criterion **C** and **B** respectively, leaving no additional criteria to be included.

	Providing lesson input				
EXP/Lit-->	**A**	**B**	**C**	**D**	**E**
1	**10.0**	3.7	3.7	1.0	5.7
2	1.3	4.7	**7.0**	3.0	5.0
3	2.0	**8.7**	**6.0**	2.7	**8.7**
4	4.3	2.7	3.3	2.7	2.0
5	1.0	4.7	5.0	4.3	5.7

Presented in this table, are the averages of the goodness of match analysis as conducted by our expert team for the competency: “Encouraging speaking and writing”. The Expert (Down) and literature (right) criteria are, for clarity reasons, replaced by boldly printed letters and numbers respectively. Original criteria can be found in appendix (nr). Significant matches (6 or higher) are printed bold and red.

With regards to the competency:” **Using projects for CLIL**”

When we look at the analysis of the data as provided by our expert team we see that the input from our expert team has not provided validation for literature criterion **A and B (A**: “*Uses projects so that learners learn something new and are challenged”*. **B**: “*Uses projects where learners have to think and ‘transform’ information from one form into another”.* Literature criterion**D:**“Uses projects which involve cooperative learning*”*.) shows a marginal match (<6 but >7) with expert criterion **4** (“*Teacher makes groups suitable for cooperative learning.”*) Literature criterion **C** (“*Uses projects which take learner differences into account”*)has a marginal match (>6 and <7) with expert criterion **1** (“*Teacher creates projects in which language and content skills are implemented”*). Literature criterion **E** (”*Uses projects in which both language and content are assessed)*also shows a match (>7) with expert criterion **1.** The above implies that only literature criterion **C**, **D** and **E** have been validated by our goodness of match analysis and there has been not support to include any subsequent expert criteria.

Using projects for CLIL					
EXP/LIT-->	**A**	**B**	**C**	**D**	**E**
1	4.7	5.0	**6.5**	4.0	**7.0**
2	4.0	5.0	4.3	3.3	2.7
3	1.0	1.0	3.3	2.3	2.7
4	1.0	1.0	4.0	**6.3**	1.0

Presented in this table, are the averages of the goodness of match analysis as conducted by our expert team for the competency: “Using projects in CLIL”. The Expert (Down) and literature (right) criteria are, for clarity reasons, replaced by boldly printed letters and numbers respectively. Original criteria can be found in appendix (nr). Significant matches (6 or higher) are printed bold and red.

With regards to the competency:”**Guiding and understanding**”

When we look at the analysis of the data as provided by our expert team we see that literature criterion **A** (“*Uses a variety of activities and scaffolds learning to help learners deal with input.”)* has a strong match (>8) with expert criteria **1** (“*Scaffolds students learning via a variety of explanations”)*. Literature criteria **B** (“*Points out language features of input, e.g.: text type, typical aspects of grammar, vocabulary”)* shows a marginal match (>6 but <7) with expert criterion **4** (*“Provides multimodal ways of dealing with language and content problems”)*. Literature criterion **C** (“Engages all the learners all the time in English.”) shows a near perfect match (>9) with expert criterion **3** (“*Makes/lets students talk in English*”). Literature criterion **D** (“*Asks questions which challenge the learners to think and encourage output.”)* shows a marginal match (>6 but <7) with expert criterion **1** and a near perfect match with expert criterion **5** *(“Teachers asks relevant and challenging question”).* Our goodness of match analysis did not provide validation for literature criterion **E** (“*Teaches and recycles vocabulary actively and multimodally.*”*)* The above implies that literature criteria **A, B, C** and **D** all have been validated and should be incorporated in the final rubric. Expert criteria **1** is not added as an additional criteria as it has been found to be largely redundant with the already included literature criteria **A.**

	Guiding and understanding				
EXP/Lit-- >	**A**	**B**	**C**	**D**	**E**
1	**8.7**	3.3	3.7	**6.0**	4.3
2	4.5	3.5	4.5	5.5	3.5
3	2.3	3.0	**9.3**	3.3	4.3
4	4.0	**6.0**	4.0	3.7	4.3
5	5.7	3.7	5.0	**9.3**	4.3

Presented in this table, are the averages of the goodness of match analysis as conducted by our expert team for the competency: “Guiding and understanding”. The Expert (Down) and literature (right) criteria are, for clarity reasons, replaced by boldly printed letters and numbers respectively. Original criteria can be found in appendix (nr). Significant matches (6 or higher) are printed bold and red.

With regards to the competency:”**Assessing learning**”

When we look at the analysis of the data as provided by our expert team we see that literature criterion **A** (“*Gives test or other tasks which are clearly aligned with the content and language taught.”)* has a strong match (>8) with expert criteria **1** (“*Teacher uses a range of assessment techniques for language and content aims.”)*. Literature criteria **B** (“*Gives on-the-spot and delayed feedback on content, spoken and written language in a variety of ways”)* shows marginal matches (>6 but <7) with expert criteria **3** and **4** (*“Teacher actively corrects and compliments on language use of students.”* And *“Provides feedback (unspecified) on written content*“respectively). Furthermore, there was a good match (>7) with expert criterion **1.** Literature criteria **C** was not validated by our goodness of match analysis. Literature criteria **D** and **E** (“*Assesses learning in a variety of ways e.g.: pen-and-paper tests, projects, spoken* *interaction, using visuals*” and “*Uses peer and self-assessment*” respectively) both matched with expert criteria **1** and **2**(*“Teacher uses peer assessment”*). However, literature criteria **D** showed a good match with expert criteria **1** and a marginal match with expert criteria **2**. These relations were reversed for literature criteria **E,** showing a perfect match witch expert criteria **2** and a marginal match with expert criteria **1.** The above implies that literature criteria **A, B, D** and**E** have been validated by our goodness of match analysis. Literature criterion **C** has not been validated and thus will not be included in the final rubric. Expert Criteria **3** and **4** have been deemed largely redundant with literature criteria **4** and will thus also not be added to the final rubric.

Assessing learning					
EXP/Lit-->	**A**	**B**	**C**	**D**	**E**
1	**8.0**	**7.3**	5.0	**7.7**	**6.0**
2	3.3	3.0	3.3	**6.3**	**10.0**
3	1.7	**6.3**	1.0	4.0	1.0
4	2.7	**6.7**	2.7	4.3	2.0

Presented in this table, are the averages of the goodness of match analysis as conducted by our expert team for the competency: “Guiding and understanding”. The Expert (Down) and literature (right) criteria are, for clarity reasons, replaced by boldly printed letters and numbers respectively. Original criteria can be found in appendix (nr). Significant matches (6 or higher) are printed bold and red.

Appendix 2 (Final rubric with validated criteria)

Activating prior Knowledge

The teacher	Insufficient	Sufficient	Good	Excellent
Helps Learners to activate their existing subject knowledge and experience in a variety of ways				
Helps Learners to activate their existing language knowledge and skills in a variety of ways				
Encourages learners really tot think about a new topic				
Creates activating activities which appeal to different learning styles or multiple intelligences				
Checks availability of necessary vocabulary				

Using project for CLIL

The teacher	Insufficient	Sufficient	Good	Excellent
Uses projects which take learner differences into account.				
Uses projects which involve cooperative learning.				
Uses projects in which both language and content are assessed.				

Assessing learning

The teacher	Insufficient	Sufficient	Good	Excellent
Gives test or other tasks which are clearly aligned with the content and language taught				
Gives on/the/spot and delayed feedback on content, spoken and written language in a variety of ways				
Assesses learning in avariety of ways (pen/and/paper test, projects, spoken interaction, using visuals)				
Uses peer and self-assessment				

Guiding and understanding

The teacher	Insufficient	Sufficient	Good	Excellent
Uses a variety of activities and scaffolds learning to help learners deal with input				
Points out language features of input (e.g. text type, typical aspects of grammar, vocabulary)				
Engages all the learners all the time in English				
Asks questions which challenge the learners to think and encourage output.				

Encouraging speaking and writing

The teacher	Insufficient	Sufficient	Good	Excellent
Sets up authentic writing tasks which are authentic text types including a purpose, an aim and an audience				
Stimulates and challenges learners to produce more complex output				
Teachers uses activating work formats (e.g. authentic role play)				

Providing lesson input

The teacher	Insufficient	Sufficient	Good	Excellent
Provides multimodal input (e.g. visuals, dvds, diagrams, models)				
Provides input at the appropriate language and cognitive level				
Helps learners notice and understand language features of different types of input				
Adjusts their own language to their learners (up or down)				

Appendix 3 (Tables of comparison: showing literature criteria and criteria created by our expert group of TTO teachers)

Literature → Expert ↓	Helps learners to activate their existing **subject** knowledge and experience in a variety of ways.	Helps learners to activate their existing **language** knowledge and skills in a variety of ways.	Encourages learners really to think about a new topic.	Creates activating activities which appeal to different learning styles or multiple intelligences.	Stimulates interaction between learners so that they communicate with each other.
Introduces New Topic (content). Links to what is already known and provides context /connects to other topics					
Visual representation of collective knowledge					
Engages students in various ways with focus on language/speaking: e.g. via Brainstorm					
Checks availability of necessary vocabulary					

<u>Activating Prior Knowledge: At the activating stage of a lesson, a CLIL teacher:</u>

Providing lesson input for CLIL: When providing lesson input, a CLIL teacher:

Literature → Expert ↓	Provides multimodal input (e.g. visuals, dvds, diagrams, models).	Provides input at the appropriate language and cognitive level	Helps learners notice and understand language features of different types of input	States language as well as content aims.	Adjusts their own language according to their learners (up or down).
Uses multimodal input . (e.g. Pictures/Board/Clips and videos)					
Focuses on helping students to understand input . Attention for used grammar and vocab.					
Language used by teacher is clear. Teacher Has wide vocabulary range and provides/uses necessary jargon					
Clear guidelines when using multimodal input					
Enages students					

Guiding understanding for CLIL: When guiding understanding, a CLIL teacher:

Literature → Expert ↓	Uses a variety of activities and scaffolds learning to help learners deal with input.	Points out language features of input (e.g. text type, typical aspects of grammar, vocabulary).	Engages all the learners all the time in English.	Asks questions which challenge the learners to think and encourage output.	Teaches and recycles vocabulary actively and multimodally.
Scaffolds students learning in a variety of ways .					
Uses a variety of explanations.					
Makes/lets students talk in English					
Provides multimodal ways of dealing with language and content problems					
Teachers asks relevant and challenging question					

Encouraging speaking and writing in CLIL: When working on speaking and writing, a CLIL teacher:

Literature → Expert ↓	Provides learners with plenty of speaking and writing opportunities, ensuring that learners always use English.	Sets up pair and group work where learners communicate in English.	Sets up authentic writing tasks which are authentic text types including a purpose, an aim and an audience.	Scaffold output e.g. by providing learners with speaking or writing frames.	Stimulates and challenges learners to produce more complex output
Teacher actively promotes and checks interaction with peers and teacher in English					
Teacher sets up moments for student presentations					
Teacher makes writing task (language/content) integral part of class					
Teacher allows for additional time for answering question					
Teacher uses activating work formats (authentic role-play)					

Assessing Learning

Literature → Expert ↓	Gives test or other tasks which are clearly aligned with the content and language taught	Gives on-the-spot and delayed feedback on content, spoken and written language in a variety of ways	Makes assessments criteria clear to learners	Assesses learning in a variety of ways (pen-and-paper tests, projects, spoken interaction, using visuals)	Uses peer and self assessment
Teacher uses a range of assessment techniques for language and content aims.					
Teacher uses peer assessment					
Teacher actively corrects and compliments on language use of students.					
Provides feedback (unspecified) on written content .					

Using projects for CLIL: When working with projects, a CLIL teacher:

Literature → Expert ↓	Uses projects so that learners learn something new and are challenged	Uses projects where learners have to think and 'transform' information from one form into another	Uses projects which take learner differences into account	Uses projects which involve cooperative learning	Uses projects in which both language and content are assessed
Teacher creates projects in which "various skills" are implemented (language/ content?)					
Teacher creates projects with visual output					
Teacher provides individual feedback on project succes					
Teacher makes groups suitable for cooperative learning.					

Appendix 4 (Example of descriptors designed by the research team)

Providing lesson input for CLIL: When providing lesson input, a CLIL teacher:

	Insufficient	Sufficient	Good	Excellent	**(Extra)**
Provides multimodal input (e.g. visuals, dvds, diagrams, models).	The teacher does not use any multimodal input.	The teacher uses at least one form of multimodal input.	The teacher provides more than one type of multimodal input and reaches more than 50% of the learners.	The teacher provides more than two types of multimodel input and reaches more than 75% of the learners.	**The teacher provides more than two types of multimodel input and reaches more than 50% of the learners.**
Provides input at the appropriate language	The teacher provides input that is not at	The teacher provides input at the language and/or cognitive	The teacher provides input at the language and /or cognitive	Provides input at the appropriate language and cognitive	**The teacher provides input at the appropriate language and**

and cognitive level.	the appropriate language and / or cognitive level.	level that is understandable to 40 % of the learners.	level that is understandable to 50% of the learners.	level that is understandable for more than 80% of the learners.	**cognitive level that is understandable for all, but reaches only 50% of the learners.**
Helps learners notice and understand language features of different types of input.	The teacher never helps learners to notice and understand language features of different types of input.	The teacher helps learners to notice and understand language features of different types of input but doesn’t reach more than 25% of the learners.	Helps learners helps learners to notice and understand language features of different types of input and reaches 50% of the learners.	Always helps learners notice and understand language features of different types of input and reaches more than 80% of the learners.	**The teacher uses different methods to help learners notice and understand language features, and reaches more than 75%.**
States language as well as content aims.	The teacher does not state language aims at all.	The teacher focuses for 95% on content aims and for 5 % of language aims.	The teacher focuses for 90% on content aims and for 10 % of language aims.	The teacher focuses for 85% on content aims and for 15 % of language aims.	**The teacher focuses for 80 % on content aims and for 20 % of language aims.**
Adjusts their own language according to their learners (up or down).	The teacher does not adjust their language level to learners.	Adjusts their own language level to their learners in ¼ of his lessons.	The teacher adjusts their language level to their learners in 2/4 of his lessons.	The teacher adjusts their language level to the learners in ¾ of his lessons.	The teacher adjusts their own language according to their learners in 2/4 of his lessons but always tries to encourage the learners to go to a higher level.