DIDACTICAE | Universitat de Barcelona | ISSN 2462-2737 | DOI: 10.1344/did.2017.2.43-58

MONOGRAPH
Didactics of Physical Education
New topics, new contexts

DESIGN AND VALIDATION OF A TOOL TO EVALUATE PHYSICAL EDUCATION AND LANGUAGE INTEGRATED LEARNING TASKS

Recepción: 02/02/2017 | Revisión: 29/03/2017 | Aceptación: 30/04/2017

Josep CORAL
Universitat Autònoma de Barcelona
Josep.Coral@uab.cat

Gerard ESQUERDA
Escola Guerau de Peguera
gesquerd@xtec.cat

Judit BENITO
Escola Miralletes
jbenito5@xtec.cat

Abstract: This article describes the design and validation of a tool to evaluate physical education (PE) learning tasks in content and language integrated learning (CLIL) context. It is detailed how the tool was developed and then how its reliability and validity were tested by means of a five-phase validation process. A preliminary content validity test of the tool was conducted by five expert judges and yielded an inter-rater agreement of 89% and kappa index of 0.82. Subsequently, pilot testing of the tool was conducted and Cronbach's alpha analysis was applied to the resulting data and it showed an internal consistency of 0.84. These results suggest that the instrument described here is indeed valid to evaluate CLIL tasks.

Keywords: physical education; CLIL tasks; evaluation tool; validation study; CLIL teacher training.

DISEÑO Y VALIDACIÓN DE UN INSTRUMENTO PARA EVALUAR LAS TAREAS DE APRENDIZAJE INTEGRADO DE EDUCACIÓN FÍSICA Y LENGUA EXTRANJERA

Resumen: *El artículo describe el diseño y validación de un instrumento para evaluar tareas de educación física (EF) en un entorno de aprendizaje integrado de contenidos y lengua extranjera (AICLE). Se detalla el desarrollo del instrumento y el proceso de cinco fases usado para su asegurar su fiabilidad y validez. Se llevó a cabo una prueba preliminar de validez de contenido mediante la participación de cinco jueces expertos arrojando un porcentaje de acuerdo del 89% y un índice Kappa de 0'82. Posteriormente, se realizó una prueba piloto a cuyos resultados se le aplicó Alpha de Cronbach mostrando una consistencia interna de 0'84. Los resultados obtenidos sugieren que el instrumento descrito en el presente artículo es válido para evaluar actividades de educación física AICLE.*

Palabras claves: *educación física (EF); tareas AICLE; instrumento de evaluación; proceso de validación; formación del profesorado AICLE.*

Introduction

This paper is concerned with the process by which tools to evaluate CLIL tasks are designed and validated. It is commonly accepted that tasks are an important component in teaching since they are at the heart of teachers' daily practice, planning and decision-making processes, and a considerable amount of literature has therefore been devoted to the topic. To begin with definitions in PE, the notion of task has been widely discussed (Famose, 1992; Florence, 2008; Parlebas, 2008; Siedentop, 1998). According to Siedentop (1998), a task is composed of a goal and a series of steps that are needed to reach it. He identified three task systems in PE (managerial, transitional and instructional) while Famose (1992) understands the task as the basic unit in motor learning and classifies it in clearly defined, semi-defined and undefined. By contrast, and in the context of language learning, Nunan (1991) defines tasks in terms of the curricular goals they are intended to serve; they include the input and activities that lead to the desired output. Willis (1998) considers a task as a goal-oriented activity which is directed toward a clear purpose and presents a task-based learning framework based in three components: pre-task, task-cycle and language focus. Other studies related to foreign language teaching and CLIL have included definitions from different sources. This is the case of Llinares & Dalton-Puffer (2015:77), who argue that 'the four basic criteria for tasks as defined by Ellis (2003) for task-based learning also hold for tasks as implemented by subject educators in CLIL classrooms', though they conclude that CLIL and task-based learning tasks are different in nature. Ellis (2003) is also cited in Nikula's (2015) analysis of the balance between content and language in hands-on tasks in CLIL physics and chemistry. The idea of the task as the core of instruction is also referred to in Coral and Lleixà (2013), who demonstrate the effectiveness of a PE-in-CLIL programme through task analysis, as well as in a study by Tomlinson and Masuhara (2009) on the benefits of playing physical games for language learners.

1. Designing and evaluating CLIL tasks

When it comes to auditing CLIL tasks, the CLIL matrix (Coyle, Hood & Marsh, 2010) is widely used since it offers the possibility of rating CLIL tasks according to the cognitive and linguistic levels of the different tasks of a unit. From the perspective of CLIL materials, a relation exists between the lack of commercially produced materials (or published but not suitable for teachers' needs and context) and the fact that teachers tend to use and share self-made teaching materials (Morton, 2013). Others have been designed by expert bilingual teachers in collaboration with researchers (Moore & Lorenzo, 2015; Coral, 2013) and published on-line by educational administrations . To design materials, Mehisto (2012) identifies ten criteria for the development of quality CLIL materials which take into account the added challenges that CLIL involves. For his part, Meyer (2010) refers to the still limited methodological resources and practical guidance available to teachers who wish to teach in CLIL settings. He notes the need to design and develop an instrument to analyse CLIL tasks, emphasizing the fact that CLIL teachers need tools that can help them plan, adapt or create their own materials. To design these tasks, teachers should bear in mind that classroom materials must imply rich input, must connect the real world with students' daily lives to be meaningful, and must be presented in different ways using a wide range of materials and platforms in order to be challenging. Yet at the same time classroom materials also need to

be authentic, meaning that they must provide suitable fixed scaffolding (in the form of flashcards, word cards, sentence formation charts, etc.) and dynamic scaffolding (e.g., teacher feedback). Teachers are thus urged to use or design scaffolding that lets students understand all kinds of language input, but also to make sure that the input provided is as rich as possible. Another important issue is the promotion of interaction and student production because «student interaction and output is triggered by tasks, which is why task design is at the heart of every CLIL lesson and one of the key competences for every CLIL teacher» (Meyer, 2010:17). By the same token, a well-designed task must include the intercultural dimension in order to enable students to understand a globalized world, because ultimately they may well end up working in international teams with co-workers from different cultures and nationalities, but also, more immediately, because they need to know how to interact appropriately with other cultures, each of which has its own hidden codes. Meyer also advocates enhancing the high order thinking skills (HOTS) and reflecting carefully on the use of the language when designing CLIL tasks. If properly designed, the various classroom strategies carried out by teachers can lead to real student-to-student interaction, authentic communication and subject-specific study skills. The more scaffolding students receive, the more output will flow, yielding significant improvements in basic interpersonal communication skills as well as cognitive academic language proficiency. As Meyer notes, «passive knowledge has to be turned into active knowledge» (2010:22).

Classroom teaching skills are also a key point to evaluate in teacher training and subsequent professional learning. CLIL teacher trainers use a variety of tools to evaluate CLIL teaching. Some of them have been taken from the immersion classroom and adapted to CLIL settings such as the Planning and Observation Checklist (Mehisto, Marsh & Frigols, 2008), the effective CLIL Teaching Observation Tool (De Graaff, Koopman, Anikina & Westhoff, 2007) and the SIOP Model Lesson Observation Protocol (Echevarría, Voght & Short, 2010). Others have been custom-designed for CLIL, most of them compiled in Florit's (2010) study on effective teaching in CLIL. Regarding tertiary education, Sagasta & Ipiña (2016) recently developed a tool to analyse CLIL units of work in the teacher education programme at Mondragon University in Spain. In the field of PE, Rink (2013) presents the research focused on teacher effectiveness in physical education and the instruments to observe and evaluate PE teaching emphasising «the importance of developing and establishing valid and reliable tools and processes of evaluation for the field of physical education» (Rink, 2013:417).

1.1. Validation of teaching-related tools

Various methods have been used to validate teaching-related assessment instruments. For example, we find validation instruments addressed specifically to physical activity such as the systematic observation instrument for teaching games in Physical Education (PE) developed by Roberts & Fairclough (2012), who use a five-stage system: observer training, expert consultation, primary pilot testing, the training of an observer unfamiliar with the instrument and confirmation of the truthfulness of the instrument through the comments and feedback of PE specialists. Another study by Pérez-Cañado & Ráez-Padilla (2012) focuses on the design and validation of a questionnaire aimed at analysing the functioning of the European Credit Transfer System. In this case, after des-

cribing the questionnaire's features and the procedure used to code the items, they present a two–phase validation process based on expert judgements and a pilot study. A different questionnaire, also related to higher education, is validated in the work of Visser-Wijnveen, Stes & Van Petegem (2012) which aimed to clarify teachers' motivation with respect to their teaching in university settings. A compilation based on three existing questionnaires, their instrument is scored on a Likert scale. A further Likert scale questionnaire intended to determine how specialist teachers at the primary and secondary levels add key competences to their PE teaching programmes has been validated by Lleixà, Capllonch and González (2015), also by means of expert judgements and a pilot study. In language learning, specific tools have been created and validated to learn how teachers can help students in the writing process (Kear, Coffman, McKenna & Ambrosio, 2000) and writing competency (Daly, 1978). Norman and Calfee developed a test focused on infant education to «quickly evaluate early readers' and writers' understanding of letters, sounds, words and sentences» (2004:43). In a much earlier study, Harro (1997) validated a kindergarten-level physical activity questionnaire to gather parental and teachers' reports using two objective tools of physical activity assessment as reference measures of validation. A completely different approach was followed by Filardo, González-Cascos & Riesco (2011) when they evaluated the validity of systemic functional grammar as a tool for choosing classroom materials in CLIL settings. The tool presented by the authors was applied to third and fourth year primary-level science textbooks to prove that text analysis can be useful to select CLIL classroom material. Others assessed existing tools as Rowe, Schuldheisz & Van der Mars (1997) checking the validity of the scale used in the System for Observing Fitness Instruction for specifically measuring children's physical activity in elementary and middle school. Finally, Erdogan, Özel, Uşak & Prokob (2009) used an eight-step process to validate an instrument to measure university students' attitudes towards biotechnology. Taken as a whole, these studies show that there are a wide range of validation instruments available in the area of interest here. Nevertheless, specific areas of teaching still lack validated tools, and we will endeavour to fill one such gap by means of the present study.

2. Objective and research questions

Since, to our knowledge, no study has yet validated and published a specific tool for evaluating PE-in-CLIL tasks, the objective of this study is to design and validate an easy-to-use tool to evaluate PE-in-CLIL tasks specifically aimed at PE teachers in primary and secondary education as well as CLIL teacher trainers. In particular, the study will explore the following research questions:

RQ1: What variables and indicators can be used to evaluate PE-in- CLIL tasks?

RQ 2: Considering common validation procedures, what procedure can be applied in the context of this study?

RQ 3: Once applied, does the proposed procedure confirm the reliability and validity of the PE-in-CLIL task-evaluation tool presented here?

We answer these RQs by:

- Describing the context where the study took place.
- Describing the process used to establish variables and indicators.

- Identifying a suitable procedure to validate the instrument.
- Analysing the results of the validation process.

3. Method

3.1. Context

This study is part of an Action Research (AR) project carried out during the 2014-2016 school years with the support of the Catalan government's Department of Education and the Faculty of Education of the Autonomous University of Barcelona. The goal of the AR project was to provide teachers with effective, evidence-based and inquiry-based knowledge in order to successfully negotiate PE-in-CLIL lessons and language-oriented physical games at their schools (Coral & Lleixà, 2017) within a competences-based curriculum. By emphasizing that teaching PE through CLIL approach is just another way to mobilise 'practical skills, knowledge, motivation, ethical values, attitudes, emotions and other social components and behaviour' (Lleixà, González-Arévalo & Braz-Vieira, 2016. This AR project is based on the works of Casey, Dyson & Campbell (2009), Coral & Lleixà (2016), Elliot (1991, 2005, 2007) and Zwozdiak-Myers (2012), who understand AR as a key tool for teachers professional learning. Like all educational AR projects, it seeks the improvement in the practice of teaching through reflection and research. The project received essential support from the Specific Educational Resource Centre for Innovation and Educational Research (CESIRE), a unit of the Catalan Department of Education created in late 2014. One of its aims is to closely monitor research in teaching and education from schools, universities and other institutions so that the results can be promoted and adapted to meet teachers' needs. The validation tool that is presented in this paper is one of the outcomes of the AR project.

3.2. Development process of the evaluation tool for CLIL tasks

A five-phase model was used to develop the evaluation tool: Review of the literature, development of the tool, content validity of the tool, pilot testing, internal consistency of the tool.

3.2.1. Phase one: Review of the literature

To ensure that a validated tool to evaluate PE-in-CLIL tasks has not been published yet, a search of 11 education and humanities data bases (Cambridge Journals Online, Connexions, Dialnet, ERIC, HKJO, ISOC, Oxford Journals, SciELO, Science direct Teacher's Reference, Web of Science) was done using the keywords 'CLIL evaluation tasks' and 'evaluación tareas AICLE' and yielded 115 entries, although none of them proved to be related to the design and validation of a specific tool to evaluate tasks in CLIL contexts, PE included. This confirmed the gap in the CLIL body of knowledge and provided justification for the development of an easy-to-use tool to evaluate PE-in-CLIL tasks. From the entries, the existing literature on CLIL and PE-in-CLIL materials, teaching observation tools and validation procedures used in the field of education was then carefully reviewed. All those works that were deemed to be of particular relevance to this study are discussed above, in the introduction section of this article.

3.2.2. Phase two: Development of the tool

In defining what variables and indicators would best index the quality of a PE-in-CLIL teacher-

Coral, J., Esquerda, G., & Benito, J. (2017). Design and validation of a tool to evaluate physical education and language integrated learning tasks. *Didacticae*, 2, 43-58.

designed task, an essential source is the work of Meyer, who lays the foundations of a planning tool called the «CLIL Pyramid» (2010:24) in six quality principles or strategies.

Variable 1: Rich input
Indicators:

1.1 The task is meaningful since it deals with real problems.

1.2 The task connects with children's areas of interest.

1.3 Authentic language input is used to present and execute the task.

Variable 2: Scaffolded learning
Indicators:

2.1 Scaffolding facilitates and helps students understand the content and language.

2.2 The scaffolding enables students to accomplish the task through supportive structuring.

2.3 The scaffolding also supports language production.

Variable 3: Rich interaction and pushed output
Indicators:

3.1 The task provides opportunities to transfer a lot of information among students.

3.2 The task proposes situations where students are asked to interact using the language.

Variable 4: Adding the (Inter) cultural Dimension
Indicators:

4.1 The task contains differentiation strategies to accommodate all students' needs.

4.2 The task promotes personal and social competences. respecting and taking into account intercultural communication.

Variable 5: Make it HOT
Indicators:

5.1 The task creates an environment in which students are engaged and challenged with various types of thinking (LOTS and HOTS).

5.2 The task includes any type of language scaffolding to facilitate the verbalisation of thinking skills.

Variable 6: Sustainable learning
Indicators:

6.1 The task promotes connections between previous and new knowledge.

6.2 The task progression is clear and well structured.

6.3 The language activities included in the task are consistent with the lexical approach.

Given the analysis and arguments that he presents (Meyer, 2010:13-22) and the fact that the CLIL Pyramid has been applied in pre- and in-service teacher training courses across Europe, it was decided that the above-mentioned variables and indicators were appropriate to our own purposes. The tool was then prepared in a spreadsheet to be used in the validation process.

3.2.3. Phase three: Content validity of the tool

The content validity of the tool was rated by a group of experts consisting of four trained and

experienced CLIL teachers from four different schools and three subjects (PE, Music and English as a Foreign Language) and one academic, PE and CLIL expert and teacher trainer. Each of these five experts judged the relevance of the inclusion of these indicators in a sample CLIL task chosen randomly from among 20 tasks that were available in the first AR year. A four-point Likert scale (1 = not at all relevant, 2 = slightly relevant, 3 = moderately relevant and 4 = very relevant) was used. In order to identify the indicators judged insufficiently relevant, inter-judge reliability was first checked by applying the formula proposed by House, House and Campbell (1981:37-57) and then calculating the Kappa statistic.

According to House et al. (1981), there is considerable consensus that an average of agreement at or above 70% is necessary in order to show that raters are consistent in their judgements among themselves. We decided to be slightly more demanding, so that when we found less than 75% agreement for any given indicator, further clarification was necessary and the judges would be asked to revise their judgements in the hope of obtaining a higher kappa value. Like most correlation statistics, kappa can range from -1 to +1. We use Altman's interpretation (1991), which has been widely used in educational research (Torres and Peguera, 2009): K values between 0.81 and 1.00 indicated very good agreement, between 0.61 and 0.80 good, 0.41 and 0.60 moderate, while values between 0.21 and 0.40 indicated fair agreement and below 0.20 was poor. In earlier research measuring agreement among observers, Landis and Koch (1977) regarded a range between 0.61 and 0.80 as indicating substantial agreement and any rating above 0.81 almost perfect, pointing out that «although these divisions are clearly arbitrary, they do provide useful benchmarks for the discussion» (1977:165).

3.2.4. Phase four: Pilot testing

The pilot testing involved a natural sample of 20 teachers who had taken part successfully in two PE-in-CLIL training courses during 2015 and 2016. 60% of these participants were involved in primary education, 30% in secondary education and 10% in professional training programmes. They evaluated a total of 30 CLIL tasks from PE (77%) and Music (10%). The remaining 13% corresponded to physical games-oriented tasks (Tomlinson and Masuhara, 2009) prepared by English teachers to be applied in their regular English as a Foreign Language lessons. All tasks were piloted by the sample teachers during the AR process and the evaluation process followed a cross-procedure, that is, each teacher evaluated tasks that had been created and tested by another teacher. Evaluation consisted of asking each teacher to rate using a four-point Likert scale (1 = not at all relevant, 2 = slightly relevant, 3 = moderately relevant and 4 = very relevant) how closely the task seemed to fulfil the indicators described in phase 2.

3.2.5. Phase five: Internal consistency of the tool

The data for task ratings collected in the pilot study were compiled in a spreadsheet and then IBM SPSS Statistics 23.0 software was used to calculate the Cronbach's alpha coefficient, which provides a measure of the internal consistency of a scale. Cronbach's alpha is expressed as a number between 0 and 1 and was used to describe the extent to which all the items of the evaluation tool measured the same concept. According to Tavakol and Dennick (2011), there is consensus among various studies that acceptable values of alpha range from 0.7 to 0.9. Since alpha is affected by the length of the test, sample size is a consideration. However, literature on this field such as Fleiss (1981, cited in Bonett 2002) suggests that a sample size of 15-20 is sufficient for valid results

4. Results

The results of inter-judge reliability analysis using the formula proposed by House et al. (1981) are illustrated in table 1. Kappa values expressed an overall agreement of 85% and a fixed marginal kappa value of 0.74. In fact, as can be seen, all scores show either 80% or 100% agreement, with the exception of indicator 2.3, which at 60% is well below the desired 75% minimum.

VARIABLES	INDICATORS	JUDGES					AVERAGE	STANDARD DEVIATION	AGREEMENTS	AGREEMENTS + DISAGREEMENTS	% OF CONCORDANCE
		A	B	C	D	E					
1: Rich input	1.1 The task is meaningful since it deals with real problems.	3	4	4	4	4	3.8	.37	4	5	80
	1.2 The task connects with children's areas of interest.	3	3	3	3	3	3	0.00	5	5	100
	1.3 Authentic language input is used to present and execute the task.	3	3	3	3	3	3	0.00	5	5	100
2: Scaffolded learning	2.1 Scaffolding facilitates and helps students understand the content and language.	3	3	3	3	3	3	0.00	5	5	100
	2.2 The scaffolding enables students to accomplish the task through supportive structuring.	3	3	3	3	3	3	0.00	5	5	100
	2.3 The scaffolding also supports language production.	4	4	4	3	3	3.6	.45	3	5	60
3: Rich interaction and pushed output	3.1 The task provides opportunities to transfer a lot of information among students.	3	3	3	3	3	3	0.00	5	5	100
	3.2 The task proposes situations where students are asked to interact using the language.	2	2	2	2	2	2	0.00	5	5	100
the (Inter) cultural Dimension	4.1 The task contains differentiation strategies to accommodate all students' needs.	4	3	3	3	3	3.2	.37	4	5	80

	4.2 The task promotes personal and social competences respecting and taking into account intercultural communication.	1	2	2	2	2	1.8	.37	4	5	80
5: Make it HOT	5.1 The task creates an environment in which students are engaged and challenged with various types of thinking (LOTS and HOTS)	2	2	2	2	2	2	0.00	5	5	100
	5.2 The task includes any type of language scaffolding to facilitate the verbalisation of thinking skills.	2	2	1	2	2	1.8	.37	4	5	80
6: Sustainable learning	6.1 The task promotes connections between previous and new knowledge.	2	2	2	2	2	2	0.00	5	5	100
	6.2 The task progression is clear and well structured.	3	3	3	3	3	3	0.00	5	5	100
	6.3 The language activities included in the task are consistent with the lexical approach.	3	3	3	3	3	3	0.00	5	5	100

Table 1. Inter-judge reliability using the percentage agreement formula given in House et al. (1981).

Following the phase three procedure described above (section 3.2.3), the rubric for indicator 2.3 was revised and rewritten by the five judges as «The scaffolding also supports oral or written language production». When the rating procedure was repeated the item now obtained an overall percentage of agreement of 89% and a fixed marginal kappa of .82. The final model of the tool following this revision is presented in table 2.

INDICATORS

1 The task is meaningful since it deals with real problems.
2 The task connects with children's areas of interest.
3 Authentic language input is used to present and execute the task.
4 Scaffolding facilitates and helps students understand the content and language.
5 The scaffolding enables students to accomplish the task through supportive structuring.
6 The scaffolding also supports oral or written language production.
7 The task provides opportunities to transfer a lot of information among students.
8 The task proposes situations where students are asked to interact using the language.
9 The task contains differentiation strategies to accommodate all students' needs.
10 The task promotes personal and social competences respecting and taking into account intercultural communication.
11 The task creates an environment in which students are engaged and challenged with various types of thinking (LOTS and HOTS)

12 The task includes any type of language scaffolding to facilitate the verbalisation of thinking skills.

13 The task promotes connections between previous and new knowledge.

14 The task progression is clear and well structured.

15 The language activities included in the task are consistent with the lexical approach.

Table 2. Final model of the tool to evaluate PE-in-CLIL tasks.

It was this final model including the reformulated indicator that was used in the subsequent pilot study in which the 20 teachers rated 30 CLIL tasks. The reliability coefficient obtained in the pilot study was α= .84, thus demonstrating very good internal consistency. Table 3 gives the averages, standard deviations, variance and alpha values related to each of 30 tasks used in the pilot study.

Task	Mean	Std. Deviation	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	2.67	.617	98.07	49.924	.366	.839
2	3.67	.488	97.07	49.352	.570	.834
3	3.80	.414	96.93	49.924	.581	.835
4	3.07	.799	97.67	45.095	.723	.824
5	3.27	.799	97.47	44.552	.779	.821
6	3.40	.507	97.33	55.667	-.321	.857
7	2.07	.799	98.67	53.238	-.033	.856
8	3.13	.640	97.60	48.114	.560	.832
9	3.73	.458	97.00	51.714	.239	.842
10	3.40	.632	97.33	49.238	.435	.837
11	2.93	.458	97.80	52.457	.125	.845
12	3.87	.352	96.87	51.981	.274	.842
13	3.87	.352	96.87	50.981	.476	.838
14	3.33	.617	97.40	47.686	.637	.830
15	3.40	.632	97.33	46.381	.779	.825
16	3.47	.516	97.27	54.352	-.148	.853
17	2.40	.632	98.33	53.095	.000	.851
18	3.13	.640	97.60	48.971	.459	.836
19	3.93	.258	96.80	52.029	.376	.841
20	3.60	.507	97.13	50.124	.434	.837
21	2.53	.640	98.20	50.171	.321	.841
22	3.33	.816	97.40	49.257	.312	.842
23	3.73	.458	97.00	50.286	.462	.837
24	3.00	.655	97.73	47.924	.567	.832
25	3.60	.507	97.13	50.695	.352	.840
26	3.87	.352	96.87	50.838	.505	.837
27	3.53	.640	97.20	49.314	.420	.837
28	3.93	.258	96.80	53.029	.106	.844
29	3.47	.640	97.27	50.638	.269	.842
30	3.60	.507	97.13	50.695	.352	.840

Table 3. Reliability statistics for the pilot study.

The number of tasks that scored below 3 in the pilot study and the corresponding percentages are detailed in table 4. This table is quite revealing with regard to two indicators, number 9 and 13. Eight out of 30 piloting teachers (26.67%) gave tasks a rating below «3» for indicator 9, and ten out of the 30 gave tasks ratings below «3» for indicator 13. However, this was because the respective indicators were either not truly pertinent to the respective tasks involved or only marginally so.

	TASK INDICATOR	Number of tasks rated below «3»	% of tasks rated below «3»
1	The task is meaningful since it deals with real problems.	0	0.00
2	The task connects with children's areas of interest.	0	0.00
3	Authentic language input is used to present and execute the task.	2	6.67
4	Scaffolding facilitates and helps students understand the content and language.	3	10.00
5	The scaffolding enables students to accomplish the task through supportive structuring.	3	10.00
6	The scaffolding also supports oral or written language production.	2	6.67
7	The task provides opportunities to transfer a lot of information among students.	4	13.33
8	The task proposes situations where students are asked to interact using the language.	4	13.33
9	The task contains differentiation strategies to accommodate all students' needs.	8	26.67
10	The task promotes personal and social competences respecting and taking into account intercultural communication.	2	6.67
11	The task creates an environment in which students are engaged and challenged with various types of thinking (LOTS and HOTS)	2	6.67
12	The task includes any type of language scaffolding to facilitate the verbalisation of thinking skills.	5	16.67
13	The task promotes connections between previous and new knowledge.	10	33.33
14	The task progression is clear and well structured.	3	10.00
15	The language activities included in the task are consistent with the lexical approach.	3	10.00

Table 4 Number and percentage of tasks that were rated below «3» for each indicator.

5. Discussion

This study was inspired by the need to provide teachers and teacher trainers with a validated tool to evaluate PE-in-CLIL tasks. Our results point to the successful accomplishment of this goal.

In order to answer RQ1, after reviewing the existing literature on CLIL evaluation tools we came to the conclusion that the six strategies developed by Meyer (2010) would be the most suitable for our own purposes (see 3.2.2). The fact that the CLIL matrix (Coyle et al., 2010) is mainly related to the cognitive and linguistic levels of classroom tasks and the example of Visser-Wijnveen et al., (2012) in modifying and adapting previous tools both support our decision to base the new tool on Meyer's framework for measuring the quality of CLIL materials.

In answer to RQ2, on the basis of previous research specifically related to validation processes, we followed a five-phase procedure involving a review of the literature; development of the tool; a test of the content validity of the tool; pilot testing; and, finally, a test of the internal consistency of the tool. As noted, previous research informed our choice of the variables and indicators to be rated for each PE-in-CLIL task, and likewise encouraged used to carry out the rating process by means of a four-point Likert scale. Such scales are commonly used in survey research to measure the degree of agreement with a particular question, and various authors such as Dimitrova, Ferrer-Wreder and Galanti (2016), Pérez-Cañado and Ráez-Padilla (2012) and Kear et al., (2000) have based their va-

lidation process on a scale of four because respondents have no option between 2 («agree slightly») and 3 («agree moderately») and must thus commit themselves to either the positive or negative pole of agreement. Thus, if a CLIL task were to receive an average rating below 3, this would suggest that raters largely felt that the indicator in question was largely not evidenced in the task. As we have seen, this occurred in our study when teachers tested our tool on prepared CLIL materials (see table 4), though in this case the nature of the tasks being rated precluded evidence of the indicators in question. Nonetheless, this kind of information will be very useful for PE-in-CLIL teachers to self-evaluate and if necessary correct tasks of their own design, just as it will be of utility for CLIL teacher trainers in guiding novice CLIL teachers.

With regard to RQ3, both content validity and reliability were confirmed, with tests yielding an inter-rater agreement percentage of 89% and kappa value of 0.82, and an alpha value of 0.84, respectively. This goes some way to address the concerns of Pérez-Cañado (2012:331), who identified inter-rater reliability as one of «the shortcomings and flaws of previous research (in CLIL)» that need to be remedied. This validation process is one of the final outcomes of a two-year inquiry-based in-service training programme that corroborates the fact that the AR process can enhance both practical and theoretical knowledge as expressed by, amongst others, McNiff & Whitehead (2006), Elliot (1991, 2007) and López-Pastor, Monjas & Manrique (2011). It also contributes to filling the so-called gap between theory and practice (Casey, Dyson & Campbell, 2009) and fosters the knowledge transfer between academics and teachers. As noted by Harris, Chisholm and Burns (2013), academics can benefit a great deal from the real-life classroom experience of teachers and, according to Galindo, Sanz & De Benito (2011), knowledge transfer in the other direction has greatly helped teachers and teacher trainers to adapt to a new reality by incorporating a practical tool that facilitates both knowledge acquisition and its subsequent transfer.

Conclusion

In this study, the content validity and internal consistency of an easy-to-use tool to evaluate PE-in-CLIL tasks specifically aimed at teachers in primary and secondary education as well as teacher trainers has been demonstrated. A confirmatory search of 11 data bases for existing literature on tasks, PE-in-CLIL materials, teaching observation tools and validation procedures in the field of education not only revealed the absence of any validated tool but also furnished us with essential guidance in designing our own implement. On the basis of our validation study conducted in the context of Catalan CLIL teacher training, the resulting instrument based on 15 indicators should prove an excellent tool to evaluate PE-in-CLIL tasks according to the quality principles for designing materials established by Meyer (2010). The instrument can be used not only to evaluate CLIL tasks (appendix 1) but also to provide clear guidelines to design such tasks and is therefore likely to prove very useful tool in teacher training. The study was based on real classroom tasks that have been developed and implemented by teachers at Catalan schools in natural contexts and PE-in-CLIL teacher training courses, thus the findings have ecological validity, at least in the Catalan context. Since the participants were overall highly motivated teachers and positively engaged in the CLIL approach, perhaps further research should be directed at testing this instrument in other contexts in order to optimise it and, if necessary, further confirm its reliability and validity.

Acknowledgements

The authors acknowledge the collaboration provided by Mr. Xavier Vinagre and Ms. Marta Hernández during the validation process and the support provided by the Specific Educational Resource Centre for Innovation and Educational Research (CESIRE), a unit of the Catalan government's Department of Education, the research group 'Language and Education' (LED) Ref. 2014S GR1190 and the project 'Pathway 2015' (ARMIF 00001).

References

Altman, D. G. (1991). *Practical statistics for medical research*. London England: Chapman and Hall.

Bonett, D. G. (2002). Sample Size Requirements for Testing and Estimating Coefficient Alpha. *Journal of Educational and Behavioral Stadistics. 27*(4), 334-340.

Casey, A., Dyson, B., & Campbell, A. (2009). Action Research in Physical Education: focusing beyond myself through cooperative learning. *Educational Action Research, 17*(3), 407-423.

Coral, J. (2013). Physical education and English integrated learning: How school teachers can develop PE-in-CLIL programmes. *Temps d'Educació, 45*, 41–65. Retrieved from http://www.raco.cat/index.php/TempsEducacio/article/view/274635

Coral, J. y Lleixà, T. (2013). Las tareas en el aprendizaje integrado de educación física y lengua extranjera (AICLE). Determinación de las características de las tareas mediante el análisis del diario de clase. *Retos. Nuevas Tendencias en Educación Física, Deporte y Recreación, 24*, 79–84.

Coral, J. & Lleixà, T. (2016). Physical education in content and language integrated learning: successful interaction between physical education and English as a foreign language. *International Journal of Bilingual Education and Bilingualism, 19*(1), 108-126. doi:10.1080/13670050.2014.977766

Coral, J. & Lleixà, T. (2017). In-service Content and Language Integrated Learning (CLIL) teachers: An Action Research Project in professional learning in Catalonia. In P. Boyd and A. Szplit (Ed.), *Teachers and teacher educators learning through inquiry: International Perspectives* (in press). Libron: Kraków.

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

De Graaff, R., Koopman, G. J., Anikina, & Westhoff, G. (2007). An Observation Tool for Effective L2 Pedagogy in Content and Language Integrated Learning (CLIL). *International Journal of Bilingual Education and Bilingualism, 10*(5), 603–624. doi:10.2167/beb462.0.

Daly, J. (1978). Writing apprehension and writing competency. *Journal of Educational Research, 72*(1), 1-10.

Dimitrova, R., Ferrer-Wreder, L. & Galanti, R.M. (2016). Pedagogical and Social Climate in School Questionnaire: Factorial Validity and Reliability of the Teacher Version. *Journal of Psychoeducational Assessment, 34*(3), 282 –288.

Echevarria, J., Vogt, M.E. & Short. D. (2010). *Making content comprehensible for English learners*. Boston, MA: Pearson.

Ellis, R. (2003). *Task-based language teaching and learning*. Oxford: OUP.

Elliot, J. (1991). *Action research for educational change*. Berkshire: Open University Press.

Elliot, J. (2005). *La investigación-acción en educación*. Madrid: Morata.

Elliot, J. (2007). *Reflecting where the action is.* London: Routledge.

Erdogan, M., Özel, M., Uşak, M., & Prokob, P. (2009). Development and validation of an instrument to measure university tudents' biotechnology attitude. *Journal of Science Education and Technology, 18*, 255-264. DOI:10.1007/s10956-009-9146-6.

Famose, J. P. (1992). *Aprendizaje motor y dificultad de la tarea*. Barcelona: Paidotribo.

Filardo, L., González-Cascos, E. & Riesco, L. (2011). On the validity of systemic functional approaches as a tool for selecting materials in CLIL context. A case study. *Porta Linguarum, 16*, 65-74.

Florence, J. (1991). *Tareas significativas en Educación Física escolar.* Barcelona: Inde.

Florit, C. (2010). *Pràctica docent efectiva AICLE. Llicència A. Departament d'Ensenyament de la Generalitat de Catalunya.* Retrieved from http://www.xtec.cat/sgfp/llicencies/200910/memories/2052/pdeaicle.pdf

Galindo, J., Sanz, P. y De Benito, J. J. (2011). La universidad ante el reto de la transferencia del conocimiento 2.0: Análisis de las herramientas digitales a disposición del gestor de transferencia. *Investigaciones Europeas de Dirección y Economía de la Empresa, 17*(3), 111-126.

Harris, M., Chisholm, C. & Burns, G. (2013). Using the knowledge transfer partnership approach in undergraduate education and practice-based training to encourage employer management. *Education + Training, 55*(2), 174-190.

Harro, M. (1997). Validation of a questionnaire to assess physical activity of children ages 4-8 Years. *Research Quaterly for Exercise and Sport, 68*(4), 259-268.

House, A. E., House, B. J. & Campbell, M. B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. *Journal of Behavioral Assessment, 3*(1), 37-57.

Kear, K., Coffman, G. A., McKenna, M. G. & Ambrosio, A. L. (2000). Measuring attitude toward writing. A new tool for teachers. *The Reading Teacher, 54*(1), 10-23.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

López-Pastor, V.M., Monjas, R. & Manrique, J.C. (2011). Fifteen years of action research as professional development: seeking more collaborative, useful and democratic systems for teachers. *Educational Action Research. 19*(1), 153-170.

Lleixà, T., Capllonch, M. & Gonzàlez, C. (2015). Competencias básicas y programación de Educación Física. Validación de un cuestionario diagnóstico. Retos. *Nuevas tendencias en Educación Física, Deporte y Recreación, 27*, 52-57.

Lleixà, T., González-Arévalo, C., & Braz-Vieira, M. (2016). Integrating key competences in school physical education programmes. *European Physical Education Review* 1-20. DOI: 10.1177/1356336X15621497.

Llinares, A., & Dalton-Puffer, C. (2015). The role of different tasks in CLIL students' use of evaluative Language. *System, 54*, 69-79.

McNiff, J. & Whitehead, J. (2006). *All you need to know about action research*. London: Sage.

Mehisto, P., Marsh, D. & Frigols, M. J. (2008). *Uncovering CLIL*. McMillan: Oxford.

Mehisto, P. (2012). Criteria for producing CLI learning material. *Encuentro, 21*, 15-33.

Meyer, O. (2010). Towards quality-CLIL: Successful planning and teaching strategies. *Pulso, 33*, 11–29.

Moore, P. & Lorenzo, F. (2015). Task-based learning and content and language integrated learning materials design: process and product. *The Language Learning Journal, 43*(3), 334-357. DOI: 10.1080/09571736.2015.1053282

Morton, T. (2013). Critically evaluating materials for CLIL: Practitioners' practices and perspectives. In J. Grey (Ed.), *Critical Perspectives on Language Teaching Materials* (pp. 111-136). London: Palgrave McMillan.

Nikula, T. (2015). Hands-on tasks in CLIL science classrooms as sites for subject-specific Language use and learning. *System, 54*, 14-27.

Norman, K. A. & Calfee, R. C. (2004). Tile Test: A hands-on approach for assessing phonics in the early grades. *The Reading Teacher, 58*(1), 48-52.

Nunan, D. (1991). Communicative tasks and the language curriculum. *TESOL Quarterly, 25*(2), 279–295. DOI:10.2307/3587464.

Parlebas, P. (2008). *Juegos, deporte y sociedad. Léxico de praxiología motriz*. Barcelona: Paidotribo.

Pérez-Cañado, M. L. (2012). CLIL research in Europe: past, present and future. *International Journal of Bilingual Education and Bilingualism, 15*(3), 315-341.

Pérez-Cañado, M.L. y Ráez-Padilla, J. (2012). Diseño y validación de cuestionarios para la valoración del sistema ECTS. *REIFOP, 15*(3), 145-179.

Rink, J. E. (2013). Measuring teacher effectiveness in physical education. *Research Quarterly for Exercise and Sport, 84*, 407–418.

Roberts, S. & Fairclough, S. (2012). A five-stage process for the development and validation of a systematic observation instrument: The system for observing the teaching of games in physical education (SOTG-PE). *European Physical Education Review, 18*(1), 97-113.

Rowe, P. J., Schuldheisz , J. M. & Van der Mars, H. (1997). Validation of SOFIT for measuring Physical Activity of First to Eight-Grade Students. *Pediatric /Exercise Science, 9*, 136-149.

Sagasta, P. & Ipiña, N. (2016). Teacher educators growing together in a professional learning community: Analysing CLIL units of work implemented in Teacher Education. In D. Lagasabaster and A. Doiz (Eds), *CLIL experiences in secondary and terciary education* (pp. 161-196). Bern: Peter Lang.

Siedentop, D. (1998). *Aprender a enseñar la educación física*. Barcelona: Inde.

Tavakol, M, & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55.

Tomlinson, B. & Masuhara, H. (2009). Playing to learn: A review of physical games in second language acquisition. *Simulation & Gaming, 40*, 645-668.

Torres, J. J., & Peguera, V. H. (2009). Cálculo de la fiabilidad y concordancia entre codificadores de un sistema de categorías para el estudio del foro online en E-Learning. *Revista de Investigación Educativa*, 27(1), 89-103.

Visser-Wijnveen, G. J., Stes, A. & Van Petegem, P. (2012). Development and validation of a questionnaire measuring teachers' motivations for teaching in higher education. *Higher Education, 64*(2), 421-436.

Willis, J. (1998). Task-based learning. What kind of adventure?. *The Language Teacher, 22*, 7. Retrieved from: http://jalt-publications.org/tlt/articles/2333-task-based-learning-what-kind-adventure.

Zwozdiak-Myers, P. (2012). *The teacher's reflective practice handbook*. London: Routledge.

PHYSICAL EDUCATION IN CLIL TASKS EVALUATION

From Coral, J., Esquerda, G. and Benito, J. (2017). Design and validation of a tool to evaluate physical education and language integrated learning tasks. *Didacticae: Journal of Research in Specific Didactics*, 2.

Date	*A teacher-made activity.* **School:** **Teacher:**	*A published activity.* **Publisher:** **Author:**

Name of the task:	**Aim of the task:**

Ages:	**Time required:**	**Equipment & resources:**

Competence-based PE Curriculum: Mark (X) which PE dimensions are connected with the task.

Primary Education				Secondary Education			
Physical activity	Healthy habits	Expression and bodily communication	Motor games and leisure time	Healthy Physical Activity	Sport	Physical activity and leisure time	Expression and bodily communication

Write three motor contents that are developed in the task:

Please circle the number that best reflects what you observe in the task

	INDICATORS	From not evident (1) to highly evident (4)			
1	The task is meaningful since it deals with real problems.	1	2	3	4
2	The task connects with children's areas of interest.	1	2	3	4
3	Authentic language input is used to present and execute the task.	1	2	3	4
4	Scaffolding facilitates and helps students understand the content and language.	1	2	3	4
5	The scaffolding enables students to accomplish the task through supportive structuring.	1	2	3	4
6	The scaffolding also supports oral or written language production.	1	2	3	4
7	The task provides opportunities to transfer a lot of information among students.	1	2	3	4
8	The task proposes situations where students are asked to interact using the language.	1	2	3	4
9	The task contains differentiation strategies to accommodate all students' needs.	1	2	3	4
10	The task promotes personal and social competences respecting and taking into account intercultural communication.	1	2	3	4
11	The task creates an environment in which students are engaged and challenged with various types of thinking (LOTS and HOTS)	1	2	3	4
12	The task includes any type of language scaffolding to facilitate the verbalisation of thinking skills.	1	2	3	4
13	The task promotes connections between previous and new knowledge.	1	2	3	4
14	The task progression is clear and well structured.	1	2	3	4
15	The language activities included in the task are consistent with the lexical approach.	1	2	3	4

A PE-in-CLIL task is said to be balanced when it has measured equivalence between motor, communication, cognition and social/personal skills. Is the task balanced? **Yes** **No**

If you have answered No, circle which skills predominate: **motor** - **communication** - **cognition** - **social/personal skills**

Comments:

Appendix 1.