Apples – Journal of Applied Language Studies
Vol. 13, 1, 2019, 133–149

Developing and using tasks for the assessment of speaking

Neus Figueras, University of Barcelona

This article provides an overview of the development and use of tasks for the assessment of speaking. It first addresses the key role of the assessment of the skill of speaking within language assessment in general and in the context of teaching and learning foreign languages. Then, it discusses how social changes and research have reshaped the way speaking is defined and operationalized, and focuses on how speaking can be assessed more validly and reliably today. The central role of tasks is also discussed, together with the importance of taking into consideration the implications and impact of different task characteristics. Some recommendations for the development of useful and meaningful assessment tasks which foster uses of assessment that contribute to learning are also proposed. The aim is to revisit due process procedures in the development of speaking assessments with a view to problematizing how to best address 21st century needs.

Keywords: assessment, language, testing, speaking, tasks

1 Introduction

Two facts can be singled out when discussing the current increased interest in the assessment of speaking by researchers and teachers alike.

On the one hand, assessment, understood in its widest possible sense (from measurement-oriented to learning-driven approaches), is increasingly used as the crucial witness of achievement(s) in education and as a tool for enhancing efficiency in schooling (Takala, Erickson, & Figueras, 2013). This has resulted in many countries setting up their own monitoring systems and developing national and/or regional exams and surveys. Increased focus on assessment has also fostered the participation in international comparative studies, in the case of foreign languages the European Survey on Language Competences ESLC (2012), and in the growing acceptance of widely used international exams, often to the detriment of local ones. These initiatives have naturally had an impact in the media, in language education policies and in the wider social context. They have attracted the attention of researchers and have had an influence on teachers' work. The reported results and their interpretation by different stakeholders have caused changes in curricula, in the number of teaching hours, and in

Corresponding author's email: nfiguera@xtec.cat
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies
University of Jyväskylä

http://apples.jyu.fi
http://dx.doi.org/10.17011/apples/urn.201903011693

methodologies, as can be observed in the latest Eurydice Report on Languages in Education (Eurydice network, 2017).

On the other hand, speaking is being given growing importance as a catalyst in language proficiency. Human oral communication is seen not only as accountable for "the projection of the self into the world" (Hughes, 2011, p. 8) but also as the crucial tool for the development of knowledge and for social advancement in an increasingly multilingual world. Language teaching professionals today, whether working in L1 contexts, in bilingual contexts or in foreign language contexts, need to do more than to help language users and language learners to be able to show social interactional abilities in informal communication (Basic interpersonal communication skills – BICS, cf. Cummins, 1999). They also need to focus on developing the ability to function in situations that may be more cognitively demanding and ask for some features of CALP (Cognitive academic language proficiency, cf. Cummins, 1999). These changes in the consideration of language proficiency – and by extension of speaking – are caused by changes in international socioeconomic contexts, and by the increase in mobility, both real and virtual, which have multiplied multilingual scenarios in all domains of communication, including education. Language users and learners today need to show varying degrees of plurilingual competence at different levels of proficiency. Awareness of these changes and of their resulting demands can be seen in macro linguistic policy recommendations (Council of Europe, 2007) and in the increased research into the languages of schooling (Schleppegrell, 2001, 2015). It can also be seen in the development and use of language corpora in language education. Language corpora flesh out actual language exponents as delivered by native and non-native speakers when performing real life tasks and other tasks in academic or non-academic situations (Friginal, Lee, Polat, & Robertson, 2017; Gablasova, Brezina, McEnery, & Boyd, 2015; Seidlhoffer, 2011).

The following sections discuss the structured development of speaking assessment(s) in the context outlined in the preceding paragraphs. They deal, in turn, with the three modules in Luoma's (2004) framework for modular specifications, which are still valid today. They cover the different strands in current research into the assessment of speaking, namely construct issues, task issues and assessment issues. The aim is to revisit due process procedures in the development of speaking assessments with a view to problematizing how to address 21st century needs. The conclusion includes some recommendations that may contribute to closing the gap between assessment operations and learning-teaching needs.

2 Construct Issues: *why* and *what* to test

The reasons for the increasing interest in the assessment of speaking and the growing demands on language users and language learners outlined in the preceding section point to the need to clearly define the *why* and the *what* of each assessment endeavour, that is, its purpose(s) and its content. Takala, Erickson, Gustafson and Figueras (2016) state these two issues as the first to address in order to achieve good and ethical practice in language assessment. Traditionally, much more attention has been paid in teacher training, in assessment manuals and in research, to the *what*, to the definition of the construct to be assessed, downplaying the fact that construct definitions should always be dependent on

the assessment purpose(s). This is probably because assessment purpose(s) is a policy issue rather than a technical or content issue. Moreover, the *why* and the *what* constitute two sides of the same coin and are interdependent. Today, however, purpose, score interpretation and use are receiving more attention in all types and contexts of assessment and are at the heart of the argument-based approach in test validation processes (Bachman & Palmer, 2010; McNamara, 2007; Kane, 2013; Pellegrino, DiBello, & Goldman, 2016).

The argument-based approach is based on Toulmin's (2003) framework for creating informal arguments, which requires that a chain of reasoning be established that is able to build a case towards a conclusion. In this case the conclusion would be to determine the plausibility and reasonableness of the score interpretations and uses of a given test or assessment. The argument-based approach to test validation begins at the onset of any assessment project, and requires that the definition of the *what* to assess and the operationalisation of the resulting construct onto test specification considers validity *a priori* (Messick, 1996). This approach clearly takes into account the purpose of the test and the expected uses and consequences of the interpretation of its scores. Considering the varied communicative needs of foreign language users and learners (see Snow & Katz, 2014, for an exploration on the assessment of language and content in L1 and in Content and Language Integrated Learning [CLIL] contexts), most existing speaking assessments will not hold up to scrutiny. In fact BICS are the most prominent features in assessment activities (Hulstijn, 2011) even at higher levels of proficiency, and CAF (complexity, accuracy, fluency) are the most common criteria in marking schemes (Revesz, Ekiert, & Torgersen, 2016).

It is time to re-examine existing theoretical models of communicative competence to which most assessment operations claim to refer. It is also necessary to clearly identify what it means to know and use a language in different contexts today (Bachman, 1990; Bachman & Palmer, 2010; Council of Europe, 2001, 2018; Hymes, 1972). The constructs selected should form the frameworks (Fulcher & Davidson, 2007) deemed most adequate to match the different purposes in the assessment of speaking in different contexts and function as a working blueprint for test design: the test specifications. Task development and the drafting of marking schemes will then need to address the tension between, on the one hand, ability and, on the other hand, ability for use as described by McNamara (1996), as cited in Harding 2014, p. 191)

> Ability for use…is more difficult to grasp, because we need to consider a range of underlying language-relevant but not language-exclusive cognitive and affective factors (including general reasoning powers, emotional states and personality factors) which are involved in performance or communicative tasks. (McNamara, 1996, p. 59)

This tension was also identified by Bachman (2007) who held it responsible for the three differing approaches to construct definition: ability focused, task focused and interaction focused. It was also addressed by Hulstijn (2011) when he challenged the notion of 'level':

> the notion of level in most second language (L2) assessment scales…is confounded with people's intellectual functioning because higher levels of LP [Language

> Proficiency] cannot be attained by people with lower intellectual, educational, occupational or leisure-time profiles. (Hulstijn, 2011, p. 229)

Harding (2014) proposes an ambitious research agenda listing a variety of research areas for what he refers to as the "reinvigoration" of communicative language testing (CLT). The agenda features adaptability at the forefront and bears directly on the assessment of speaking. Proposals in Harding's agenda also include the development of language tests which use a variety of skills and abilities (ability to accommodate, to negotiate meaning, to ascertain and deploy appropriate pragmatics…), and presents integrated tasks, tasks utilizing social networks and tapping into new literacies, and suggest the use of newly available resources such as corpora. Harding also proposes to make us of research methods such as stimulated recall to check on both the usability and the usefulness of the proposals.

The proposed organization of language competence onto modes of communication by the Common European Framework of Reference for Languages (CEFR) published in 2001 by the Council of Europe deserves some attention here. The CEFR scales and descriptors have overshadowed the presentation of language activities which are presented in the CEFR under the modes of reception, production, interaction and mediation (see graphic representation in Figure 1 below) rather than along the division of the traditional four skills (reading, listening, speaking, writing) and the language components (grammar, vocabulary, phonology). This global view of language use, including mediation and clearly relating interaction with production and reception widens the perception of speaking as an isolated skill and of how speaking is understood in many language tests. Today many oral tests still solely consist of a picture description, or of a semi structured interview, or a combination of both. If we think of a student who needs to show a B1 speaking ability, we would expect a speaking test to gather – at least - evidence of the elements in the overall descriptors for B1 in the CEFR (Table 1), and of as many of the elements in the additional performance descriptors in the CEFR (e.g. Goal Oriented Co-operation, Transactions to Obtain Goods and Services, Informal Discussion with Friends,…)

Table 1. Overall Oral Production and Spoken Interaction B1 Descriptors (Council of Europe, 2001, pp. 58, 74).

Overall Oral Production	Can reasonably fluently sustain a straightforward description of one of a variety of subjects within his/her field of interest, presenting it as a linear sequence of points
Overall Spoken Interaction	Can communicate with some confidence on familiar routine and non-routine matters related to his/her interests and professional field. Can exchange, check and confirm information, deal with less routine situations and explain why something is a problem. Can express thoughts on more abstract, cultural topics such as films, books, music etc. Can exploit a wide range of simple language to deal with most situations likely to arise whilst travelling. Can enter unprepared into conversation on familiar topics, express personal opinions and exchange information on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).

The CEFR's global view of language use also allows us to revisit the distinction between transactional and interpersonal use of language and between interpersonal language use and the use of language for the development of ideas. The recent CEFR Companion Volume (Council of Europe, 2018) includes even more scales and descriptors that may be helpful, in particular those for mediation activities and mediation strategies (and those concerning speaking in multimodal activities typical of web use such as Online interaction and discussion or Goal-oriented online transactions and collaboration).

The new scales and descriptors respond to changes in international socioeconomic scenarios already mentioned in the Introduction, which could not be foreseen when the CEFR was published in 2001. When the CEFR was initially commissioned at the Ruschlikon symposium in 1991, teaching and learning of foreign languages had its focus on tourism and professional exchanges, mostly in Europe and often not reaching beyond B2 levels. Today, the increase in educational exchanges and migration movements caused by political, economic or religious situations, have resulted in more demands for higher and more sophisticated levels of language proficiency for professional use and for social integration. Schleppegrell (2015) describes how context(s) of use have changed and evolved, becoming overtly plurilingual and expanding beyond the interpersonal and social domains to professional and educational areas.

In mediation, as the construct is understood in the CEFR Companion volume,

> the user/learner acts as a social agent who creates bridges and facilitates the construction or conveyance of meaning, sometimes within the same language, sometimes from one language to another (Council of Europe, 2018, p. 83)

The mediation scales include an *Overall mediation scale*, which can perhaps be the best starting point, and further scales grouped into three main groups:

- *Mediating a Text* (Relaying specific information in speech and in writing; Explaining data in speech and in writing; Processing text in speech and in writing; Note-Taking, Expressing a personal response to artistic text and Analysis and criticism of artistic text),

- *Mediating Communication* (Facilitating pluricultural space; Acting as an intermediary in informal situations; Facing delicate situations and disagreements),

- and *Mediating Concepts* (Facilitating collaborative interaction with peers; Collaborating to construct meaning; Managing interaction and Encouraging conceptual talk).

The new scales for mediation strategies include Strategies to simplify a text (Elaborating a dense text and Streamlining a text) and Strategies to explain a new concept (Linking to previous knowledge; Breaking down complicated information and Adapting language). Like the other scales in the CEFR, these scales can help identify teaching, learning and assessment language activities that can be developed into tasks. The relationship between the four modes of communication in the CEFR is best appreciated in Figure 1 below. The figure was already present in a draft version of the 2001 CEFR and is now published in the Key Aspects Section in the CEFR Companion Volume (Council of Europe, 2018, p. 25). The figure presents how communication modes, macro language functions and language activities are related. Table 2 provides some examples of languages activities within each communication mode.

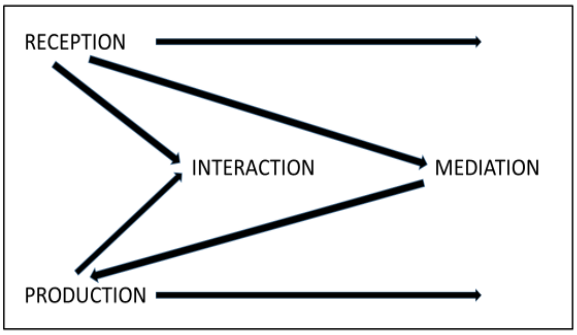

Figure 1. The relationship between reception, production, interaction and mediation.

Table 2. Macro-Functional basis of CEFR categories for communicative language activities.

RECEPTION	PRODUCTION	INTERACTION	MEDIATION
e.g. Reading as a leisure activity	e.g. Sustained monologue: describing, experience	e.g. Conversation	Mediating communication
e.g. Reading for information and argument	e.g. Sustained monologue: giving information	e.g. Obtaining goods and services Information Exchange	Mediating a text
(Merged with reading for info and argument)	e.g. Sustained monologue: presenting a case	e.g. Discussion	Mediating concepts

From what has been discussed so far, it seems that there is an embarrassment of riches when it comes to theoretical models of language proficiency from which to select the constructs to be assessed. Moreover, efforts need to be placed precisely in the identification and selection of assessment frameworks that are adequate, fair and relevant for the diverse purposes and varied contexts of today's assessments.

The obvious conclusions, not necessarily good news, to how to address construct issues in the development of speaking tests is therefore to bear in mind the saying "one size does not fit all". This makes it necessary to try and address the perennial challenge pointed out by Takala et al. (2016) in all assessment and testing practices, which is

> to avoid over-emphasizing more easily measured skills at the expense of competencies, such as reflection, critical analysis, and problem solving (Takala et al., 2016, p. 310)

The overview of the development and use of speaking assessment tasks that follows in the next section needs to be understood as a structure that frames the assessment of the different constructs selected as adequate in the specific contexts where the assessment will be used.

3 Task Issues: characteristics and delivery modes

There is considerable interest in the development of tasks in language education, shown by the existence of the Task Based Language Teaching International Association (TBLT), founded in 2005 as an:

> educational framework for the theory and practice of teaching second or foreign languages. Based on empirical research, TBLT adopts meaning-based, communicative tasks as the central unit for defining language learning needs, determining curriculum goals, designing activity in the (language) classroom, and assessing language competencies (From the first page of the association's website http://www.tblt.org/).

TBLT holds biannual conferences that include presentations on task design and task effects. Presentations and past conference proceedings are available online for consultation and bear witness of the growing body of data available on the development and use of tasks for educational purposes.

Research on task-based language assessment (TBLA) and also on tasks for the assessment of speaking has its origin in performance and occupational assessments dating back to the 1980s. There is a rich literature on the definition and use of tasks for assessment purposes (Brindley, 1994; Long & Norris, 2000), and on the problems and challenges that such an approach presents. TBLA is a relevant approach in the context where assessment, as described above, is understood as:

> the process of evaluating, in relation to a set of explicitly stated criteria, the quality of the communicative performances elicited from learners as part of goal directed, meaning-focused language use requiring the integration of skills and knowledge. (Brindley, 1994, p. 74)

Amongst the problems and challenges that using a TBLA approach to the assessment of speaking presents, Brindley (1994) points out reliability, validity and practicality issues, which will be addressed in the following section. Robinson (1996) adds to Brindley's challenges difficulties in task design, difficulties in task administration, often making tasks uneconomical. Robinson also mentions difficulties in achieving generalizability, especially if the tasks aim at tapping and reporting on information regarding some component(s) of a learner's language ability that might underlie the accomplishment of any number of different tasks. These difficulties appear mostly when the assessment of speaking aims at standardisation. Concrete tasks closely related to everyday situations and requiring the use of realia and/or support materials (e.g. solving airline problems, ordering pizzas on the phone, talking to the bank about an overdraft…) illustrate such difficulties. They may be very useful pedagogically and for classroom assessment but they may not be adequate in standardised tests. In such cases, different versions of the task need to be available. The different versions of the task must be comparable, not only with respect to the language they elicit but also with respect to the socio cognitive demands on the learner. Recent research on picture-based tasks has shown that this often easier said than done (Inoue, 2013).

Task design and development is a complex endeavour which requires putting theory into practice, savoir and savoir faire, in CEFR terms, and which requires balancing out decisions on both task characteristics and delivery modes.

Available research on task characteristics (Norris, Brown, & Hudson, 1998; Robinson, 2001; Skehan & Foster, 1999) points in the direction of cognitive factors (task complexity), interactive factors (task conditions) and learner factors (task difficulty). Following this, Skehan and Foster (1999) and Norris et al. (1998) coincide on grouping task characteristics into three main categories: cognitive complexity (input/output organization and input availability), communicative demand (mode/response level) and code complexity (range and number of input

sources). Moreover, Norris et al. (1998) include a useful appendix with example items and item generation notes which can provide insights onto how to get started in the delineation of areas and themes and in how to use the categorization proposed to develop prototypes which frame the task generation process.

Norris et al. (1998, pp. 151–227) display a set of speaking tasks in different areas and themes (e.g. food and dining is subdivided into themes like planning a dinner; ordering a pizza, ordering coffee and dessert, shopping at the grocery store...). A speaking task (ordering coffee and dessert) with a prompt like the following:

> After finishing your meal, the waiter brings a desert menu. Study the different options. When the waiter returns, order your choice of dessert. Listen to the available after-dinner beverage options that the waiter recites. Choose a beverage to accompany your desert choice. (Norris et al., 1998, p. 156)

Such a prompt may be varied in difficulty by making (linguistic) code, cognitive complexity or communicative demand high or low. Code difficulty can be low if the number of dessert options and beverages in the dessert menu is obvious and limited, and code difficulty can be high if the number of dessert and beverages variables presented to the test taker is increased. Cognitive complexity is low if there are limited options and visual aids to consider and also if little interaction with the waiter is required, which means that little on-line processing is needed. In contrast, cognitive complexity is high if there is a greater number of options and more interaction with the waiter is required (with the corresponding processing cognitive demands increased). Likewise, communicative demand is low if there is sufficient time for making choices and interaction is limited, giving the examinee high control, and it is high if reading and listening is required and the waiter and/or a partner participates actively (reducing the planning time for reaction and making the task multi-way).

On the basis of the above, the variables considered necessary, relevant and feasible to inform the design of speaking test tasks are the following (Figueras, 2001):

- Number and nature of interlocutors; whether the candidate has to give a presentation or interact with the examiner, or whether there is interaction between two or more candidates and whether they are of the same proficiency level or not, etc. The choice of examiner and of interlocutor is also considered a relevant variable (van Moere, 2012).
- Amount and type of input; in what form the candidate receives the input, whether it is through the medium of images or of language, whether the language input is written or oral and from one or more than one source, whether it is brief or not.
- Familiarity with information content and with the interlocutors; whether the response needs to be spontaneous and speeded with the candidate having to adjust to the communicative situation quickly or not and whether the candidates are acquainted with one another or not.
- Planning time; whether the candidate has time to prepare what needs to be said in the situation, whether there are suggestions on what to do during planning time or not, whether the candidate can take notes or whether the candidate can rehearse with other candidates.
- Interactional activity; what type of interaction is predicted, whether one party holds some information and for which purposes, whether the different parties have different, complementary or contrasting information…

- Communication goal; what the goal orientation of the situation is, whether there are different outcomes possible and whether these outcomes need not be conclusive – as in divergent tasks – or whether there is only a limited possible range of outcomes possible – as in convergent tasks.

Decisions on how best to address and combine the above variables or characteristics in a speaking assessment will normally point in the direction of developing a number of tasks with different objectives and salient characteristics which can tap and represent the construct(s) selected for assessment.

To illustrate how this can work, one could compare (cf. Table 3) two publicly available speaking tests for level B1 from two international exam boards, Cambridge[1] and The European Language Certificats (TELC)[2].

Table 3. Two publicly available speaking tests (Cambridge, TELC) compared.

Cambridge Preliminary English Test (PET)	**TELC English B1**
Task 1 Conversation with the examiner. The examiner asks questions and you give information about yourself, talk about past experiences, present job, studies, where you live, etc., and future plans.	Task 1: Social Contacts The task is to exchange personal information in order to get to know each other better. The candidates should say something about themselves and ask their partner questions to learn more about him or her. They can use the points on the task sheet for help but are not required to talk about all of them. The examiners may ask them to talk about an additional topic which is not on the task sheet.
Task 2 The examiner gives you some pictures and describes a situation to you. You have to talk to the other candidate and decide what would be best in the situation.	Task 2 Topic-Based Conversation The candidates have task sheets with different information on the same topic. First, each candidate should talk about the information on his or her task sheet. Then, the two candidates should exchange their opinions and talk to each other about their personal experience with the topic
Task 3 The examiner gives you a colour photograph and you have to talk about it.	
Task 4 Further discussion with the other candidate about the same topic as the task in Part 3.	Task 3 Task The task is to plan something together. The candidates are expected to exchange ideas, make suggestions and respond to the suggestions of their partner. Together, they should come up with a plan and decide who is responsible for which tasks. The points on the task sheet may be used for help.

Both tests aim at eliciting monologic and dialogic discourse (spoken production and spoken interaction respectively, as described in Table 1), and both tests also aim at tapping the contents of the descriptors in Table 1, as the different tasks require the examinees to make a description, to exchange, check and confirm information, or to express their thoughts. However, there are substantial differences in how the variables listed previously have been addressed and therefore in the task characteristics. In the Cambridge PET exam, the interlocutor is most prominent and is also the main source of input, as the visual prompts in the examinees' booklet consist of pictures and isolated words. The examinees are expected to react to the interlocutor's questions (whom they are not familiar with and who plays a dominant role) and to the tasks proposed with no planning time.

In the TELC B1 exam, the interlocutor takes the role of a facilitator, guiding the examinees into the three tasks with input given in writing in the form of instructions to complete the tasks and also in the form of linguistic and iconic prompts. The two examinees are expected to work together all the time, although they may not know one another, and this raises issues as to whether pairing of examinees needs to be random or allow the examinees' teachers or the examinees themselves to have a say. Research on the issue of pairing off examinees in an oral test is a contentious one (Ducasse & Brown, 2009) and there seems to be no unanimous recommendations about it as yet. As for planning time, whereas no clear directions have been found in Cambridge PET, the planning time the TELC B1 exam is 20 mins (with no interaction between examinees is allowed).

The differences in task characteristics pointed out so far will surely influence the examinees behaviour and their performances, as will most probably the way interactional ability is planned in the different exams. TELC B1 sets the interactional scene in a more authentic manner. In all three tasks the examinees need to talk to one another to complete them with the complementary/contrasting information they are given and not merely do as the interlocutor says, either responding to questions or reacting to prompts (cf. Cambridge PET). As for communication goals, both exams contain a divergent task (different outcomes are possible in Task 4 in the PET exam and in Task 2 in the TELC B1 exam) and a convergent task (a single outcome is expected in Task 2 in Cambridge PET "what would be best?" and in Task 3 in TELC B1 "should come up with a plan"). On the whole, TELC B1 employs a more sophisticated approach to tap the contents in Table 1, which on the one hand makes the test less dependable on the interlocutor and, on the other, makes it more authentic in terms of communicative behaviour.

However, the analyses of the speaking tests in Table 3 should not be understood as the identification of a "right" or "wrong" approach, but rather as two different ways of tackling the content, structure and organization of a speaking assessment. As already suggested in this section and also in section 2, the consideration of purpose and context needs to inform the developers on how to combine the speaking assessment variables described.

Task characteristics interact with delivery mode, whether it is live, recorded or automated, and have to be modelled following the demands of the different formats. In fact, continuous development in automated language assessments contributes interesting ideas that may affect dramatically the way the assessment of speaking will be conducted in the future. O'Sullivan (2013) presents a descriptive overview of test methods using a different format delivery (live, recorded, or automated) and a summary of their advantages and disadvantages, which test developers will find extremely useful. One must consider, however, that the overview may need to change in future as technology is moving rapidly and, as a result, oral communication styles and speaking assessment approaches and methods will be affected (van Moere, 2010).

As the process of task development evolves, the links between the constructs selected for assessment and the tasks themselves need to be documented in order to start building a validity argument *a priori*, before the test goes live (Weir, 2005). Due procedures will include consideration of how task characteristics and linguistic demands (context validity, Weir, 2005, p. 46) match the processes and resources they aim at mobilizing (theory-based validity) and are adequate to test taker characteristics. Evidence may be gathered from different sources ranging from grounded argumentation to the analysis of performances in pilot administrations or questionnaires and interviews with different stakeholders.

The evidence collected will be useful in the completion of the assessment specification, which focuses on how the test will be delivered, marked, scored, used and monitored.

4 Assessment Issues: validity and reliability

Having developed the tasks following due procedures, their use, including their scoring and rating, needs close attention in order to minimize the problems Brindley (1994) foresaw. There are a number of issues in the use of tasks for assessment purposes, namely validity, reliability and practicality. Efforts to guarantee validity and reliability need to have started during the task development process, as already stated in the preceding section. Moreover, with tasks already in place, three main areas need to be paid attention to, marking schemes, rater and interlocutor training and monitoring, and quality control. These will help guarantee scoring validity, consequential validity and criterion-related validity, the key elements in Weir's (2005, p. 46) socio-cognitive framework for validating speaking tests which need to be scrutinized once the assessments go live.

Teachers often take on the role of examiners and raters. Teachers will be making the decisions based on their interpretation of the performances through the lense(s) of a marking scheme. Their opinions and expertise need to be taken into account and training and standardization procedures need to be put in place. Validity and reliability can be increased if all parties involved, and most specially teachers and learners, are informed about the purpose and characteristics of the assessment(s), about what is expected from them, and about how they can best prepare.

On the other hand, teacher training and teacher participation in the process can have a knock on effect on the classroom and contribute to establish closer links between assessment and teaching and hence enhance learning.

4.1.Marking schemes

Scoring validity, which includes reliability in Weir (2005), depends heavily on how the tasks are administered and assessed. Whereas any speaking performance is affected by the task administrator and/or interlocutor, the relationship between a score and a given performance is always mediated by the marking scheme and by how the rater understands it (McNamara, 1996, p. 9).

Marking schemes, as pointed out by Alderson (1990), and mentioned previously, need to be developed taking into consideration the purpose of the assessment, the assessment construct, and the context where the assessment will be used, and naturally matched to the task characteristics and the delivery mode. However, the content in previous sections in this chapter suggests that special attention needs to be paid to the bands in traditionally used marking schemes, which mainly focus on complexity, accuracy and fluency (CAF) measures, to incorporate additional categories which focus on communicative adequacy and which go beyond linguistic elements to incorporate cognitive features. Revesz et al. (2016) and de Jong, Steinel, Florijn, Schoonen, & Hulstijn (2012) have researched how communicative adequacy is related to language measures and suggested changes in currently used marking schemes. Revesz et al. (2016) also pointed out the need for further research on what makes communication possible:

> The overwhelming focus on learners' lexico-grammar appears a shortcoming (Pallotti 2009), since it is well known that one can use complex and accurate language while not being functionally effective, and, vice versa, it is possible to get one's message across without using complex language and being accurate. Due to the importance of communicative success in real-world contexts, it appears timely and worthwhile to put more research emphasis on how linguistic factors may facilitate or hinder L2 users' success in completing tasks. (Revesz et al., 2016, p. 830)

4.2. Examiner (rater) and interlocutor training and monitoring

A recurrent worry in relation to the assessment of speaking is its reliability, and the standardisation both in administration procedures and interlocuting frames and in the interpretation and use of marking schemes (Brown, 2003; van Moere, 2012). Whether tasks are more or less structured, whether there is no role for an interlocutor, a fixed protocol for administration needs to be in place, including the standardisation of interlocutor and rater behaviours.

Standardisation or training needs to include a lead in phase, which allows participants to become familiar with the content and objectives of the assessment tasks, and with the type of language performances expected – both from the candidate and from the interlocutor. The initial familiarisation phase is followed by a guided practice phase where observation and analysis of exemplar and less exemplar behaviours helps understand how the content and objectives materialise, what to do and what to avoid, and allows raters to use the marking schemes and discuss their ratings with expert raters against benchmarked performances. Once accredited, both examiners and interlocutors should be monitored regularly in terms of their behaviour and discourse, and also in terms of their ratings, checking on their inter-rater and intra-rater reliability. Most international exam boards have an accreditation system of interlocutors and raters which is renewed after every few years, or increasingly every year, and many national and local exam boards have incorporated such accreditation systems. Technology allows for on line training and also for the analysis of examiner reliability, both inter-rater and intra-rater. The Into Europe Series (Csépes & Együd, 2006) published by the Hungarian Examination Board and accessible in the web offer a formidable series of videos illustrating key recommendations for standardisation behaviour and rater training.

4.3. Quality control

Quality control has become a buzzword today, and the impact of the ISO[3] Standards and the importance of auditing procedures in business has made its way onto education. Associations like EAQUALS (European Association for Quality Language Services, www.eaquals.org) or ALTE (Association of Language Testers in Europe, www.alte.org) have developed their own auditing systems and offer trained auditors to exam boards and governmental organizations. Also, assessment and testing organizations like EALTA (European Association for Language Testing and Assessment[4]) or ILTA (International Language Testing Association[5]) have published standards and codes and guidelines of good practice freely available on line which albeit not including enforcement mechanisms aim at guiding the design, development and use of language assessments. Amongst

the most valued set of available standards for the profession is the Standards for Educational and Psychological Testing published jointly by AERA/APA/NCME (2014)[6] and revised every few years. The AERA/APA/NCME Standards provide thoroughly developed criteria for the development and evaluation of tests and testing practices and guidelines for assessing the validity of interpretations of test scores for the intended test uses.

Having put together a solid assessment framework which allows for the development and administration of valid and reliable speaking assessments, there is still some further work to consider before the assessment goes live. Assessment specifications need to consider as well a protocol for *a posteriori* actions which make the test accountable and guarantee its stability over time

On the one hand, consequential validity (Weir, 2005) needs to be addressed. It is crucial to state how results will be analysed, what type of reports or feedback will be produced, for whom, and how they may or may not be used. Feedback, should be a key feature of any assessment endeavour, as it is feedback which makes assessment useful. It may have different destinataries (teachers, learners, policy makers…) and hence may take many forms, but its main purpose is to provide information on the outcomes of the assessment and contribute to improve both the assessment itself and the performances assessed.

On the other hand, criterion validity (Weir, 2005) needs to be guaranteed by plans put in place that allow for the comparability of results across sessions and with other assessments claiming similar purpose and objectives, including classroom assessments.

Striving for consequential validity and criterion validity requires the collection of evidence(s) (Pellegrino et al., 2016) and the analysis of assessment outcomes (the performances) and of results (the scores), together with the observation and analysis of the impact (the consequences) of the assessment. The nature and amount of the research into the actual behaviour of the assessment will depend on its purpose and on the resources available. But only documentation, thorough analysis and scrutiny of empirical data can provide information on the usefulness of any assessment task in tapping and assessing the construct. Such information should be the basis of the report(s) produced, which should outline grounded actions for improvement of the test itself and of its outcomes, results and consequences.

5 Conclusion and recommendations

This paper has presented recent work in the assessment of speaking, and outlined how new demands on language learners, language teachers and language testers require a broader approach and a thorough scrutiny of all the elements involved in its development and use.

If assessment is, as stated in the introduction, a crucial witness of achievement, it should be expected to help improve learning, and then what has been presented and discussed so far needs to have direct bearing in the classroom. Harlen (2006), Bachman (2010) and Pellegrino et al. (2016), amongst others, have pointed out the unresolved tension in educational assessment(s) between two poles, internal classroom assessment and external tests, often presented in opposition rather than as the two ends of a continuum. Although a lot of work has been done to give the assessments carried out in classrooms more importance, and also despite growing empirical research (Hattie, 2008; Turner & Purpura, 2015; Wiliam, 2013, amongst

many others), there is still a lot to be done to make the two poles meet and respect each other. It is to be hoped that what has been reported in this contribution finds its way into curriculum development and into the development of speaking tests in the near future. On the one hand, increased focus on context and construct relevance should imply that curricula provide context relevant, clear definitions and sufficient exponents of target language use(s) that assessment(s) can tap and relate to. On the other, the development of any assessment endeavour should incorporate a strict quality control system. Following the procedures described in the different sections, however, will not be straightforward and it will be necessary to count on all stakeholders to make it work. It has already been made explicit that teachers' expertise and knowledge can be helpful in the development and validation of marking schemes, but it is also obvious that the participation of all stakeholders is necessary in the different validation phases.

The recommendations that follow are rather obvious, and not new, although – sadly – not always paid proper attention to. A first set of recommendations has to do with what has been outlined in the first sections in the chapter, that is, the need to focus on the why and the what to assess to meet the growing need of language learners and users to be able to use language in more cognitively demanding situations in different contexts. External exam boards have to diversify their exams even more than they currently do to localise them, and to adapt them to context, purpose, age and use(s). They need to relate to different teaching programmes, different language use needs and different learning needs. Internal – school based – tests have to start tapping more sophisticated use(s) of language, and this is only possible with a thorough analysis of the curricula and the inclusion of their most relevant aspects in the assessments. In many cases, curricula may have been sufficiently defined already, and it will then be only a question of making sure that curricula and assessments match. In other cases, curricula may be undefined or underspecified, which will mean that some curriculum redrafting, revision and completion will be necessary before embarking on the development of any assessment.

A second set of recommendations has to do with the process of test development and administration. Any assessment endeavour, no matter whether it is internal or external, with a formative or summative function, having put purpose and objectives first (the *why* and the *what*), needs to follow detailed, thorough and systematic development procedures that are fit for purpose and context-relevant, and can provide valid results and feedback. This is not always the case in some external tests, which are often an example of "one size fits all" and provide, with a few exceptions, very limited feedback. Unfortunately, classroom assessments do not fare better, as they neglect thoroughness and systematicity and also offer very limited feedback. It is still very common in many classrooms that the teacher's feedback on a test or on an assessment activity is limited to a score, or to correct/incorrect marks, with no relation to curriculum objectives or suggestions for improvement.

The challenge facing those involved in the assessment of speaking is to try and use the knowledge and resources available to improve common practice(s) and already operating system(s) so that they can meet the demands of learners and society in the 21st century.

Endnotes

[1] http://www.cambridgeenglish.org/exams-and-tests/preliminary/preparation/
[2] https://www.telc.net/en/candidates/language-examinations/tests/detail/telc-english-b1.html#t=2
[3] The International Organisation for Standardisation promotes worldwide proprietary, industrial and commercial standards.
[4] www.ealta.eu.org
[5] www.iltaonline.com
[6] American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

References

AERA/APA/NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Alderson, C. (1990). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London: Modern English Publications/British Council/Macmillan.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. (2007). What is the construct? The dialectics of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L.Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 40–71). Ottawa: University of Ottawa Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Brindley, G. (1994). Task-centred assessment in language learning: the promise and the challenge. In N. Bird, P. Falvey, A. Tsui, D. Allison, & A. McNeill (Eds.), *Language and learning: papers presented at the Annual International Language in Education Conference HongKong, 1993* (pp. 73–94). Hong Kong: Hong Kong Education Department.

Brown, A. (2003). Interviewer variation and the co-construction of speaking ability. *Language Testing*, 20(1), 1–25.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2007). *From linguistic diversity to plurilingual education. Guide for the development of language policies in Europe.* Retrieved from http://www.coe.int/t/dg4/linguistic/Guide_niveau3_EN.asp#TopOfPage)

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors.* Retrieved from https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/168074a4e2

Csépes, I., & Együd, G. (2006). *INTO EUROPE The Speaking Handbook.* Retrieved from http://www.lancaster.ac.uk/fass/projects/examreform/into_europe/speaking.pdf

Cummins, J. (1999). *BICS and CALP: Clarifying the distinction*. Retrieved from https://eric.ed.gov/?id=ED438551

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5–34.

European survey of language competences. (2012). Retrieved from http://www.surveylang.org/Project-news-and-resources/Project-news-and-resources.html

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423–443.

Eurydice network. (2017). *Eurydice key data on teaching languages at school in Europe*. Retrieved from https://ec.europa.eu/education/news/20170601-eurydice-teaching-languages-school_en

Figueras, N. (2001). *Developing oral tests: can we get closer to real life?* Unpublished doctoral dissertation. University of Barcelona, Spain.

Friginal, E., Lee, J. J., Polat, B., & Robertson, A. (2017). *Exploring spoken English learner language using corpora. Learner talk*. Basingstoke: Palgrave Macmillan.

Fulcher, G., & Davidson, G. (2007). *Language testing and assessment. An advanced resource book*. New Jersey: Routledge Applied Linguistics.

Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2015). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics, 38*(5), 613–637.

Harding, L. (2014). Communicative language testing. Current issues and future research. *Language Assessment Quarterly, 11*(2), 186–197.

Harlen, W. (2006). On the relationship between assessment for formative and summative purposes. In J. A. Gardner (Ed.), *Assessment and learning* (pp. 103–119). Thousand Oaks: Sage.

Hattie, J. (2008). *Visible learning for teachers*. London: Routledge.

Hughes, R. (2011). *Teaching and researching speaking*. Harlow: Pearson Education Limited.

Hulstijn, J. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*(3), 229–249.

Hymes, D. (1972). Models of the interaction of language and social life. In J. J. Gumperz & D. Hymes (Eds.), *Directions in Sociolinguistics. The Ethnography of Communications*. (pp. 35–71).New York: Holt, Rinehart and Winston.

Inoue, Ch. (2013) *Task equivalence in speaking tests*. Bern: Peter Lang.

Kane, M. (2013) Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Long, M. H., & Norris, J. M. (2000). Task-based language teaching and assessment. In M. Byram (Ed), *Encyclopedia of language teaching and learning* (pp. 597–603). London: Routledge.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance.* London: Longman.

McNamara, T. (2007). Language assessment in foreign language education: The struggle over constructs. *Modern Language Journal, 91*(2), 280–282.

Messick, S. (1996). *Validity and washback in language testing*. Research report. Princeton: Educational Testing Service. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1996.tb01695.x/abstract

Norris, J. M., Brown, J. D., & Hudson, T. (1998). *Designing second language performance assessments*. Honolulu: University of Hawaii's Press.

O'Sullivan, B. (2013). Assessing speaking. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 156-171). New Jersey: Wiley and Sons.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81.

Revesz, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37*(6), 828–848.

Robinson, P. (1996). Task complexity and second language narrative discourse. *Language Learning, 45*(1), 99-140.

Robinson, P. (2001). *Cognition and second language instruction.* Cambridge: Cambridge University Press.

Schleppegrell, M. (2001). Linguistic features of the language of schooling. *Linguistics and Education, 12*(4), 431–459.

Schleppegrell, M. (2015). *Teaching the languages of schooling for equity and quality in education. Address to the Council of Europe Intergovernmental conference on the language dimension in all subjects: Equity and quality in education.* Strasbourg. Retrieved from

https://www.coe.int/t/dg4/linguistic/Source/LE_texts_Source/LE%202015/Schleppegrell%20Teaching%20the%20languages%20of%20schooling.pdfm

Seidlhoffer, B. (2011). *Understanding English as a lingua franca*. Oxford: Oxford University Press.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), 92–120

Snow, M. A., & Katz, A. M. (2013). Assessing language and content. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 230–247). New Jersey: Wiley and Sons.

Takala, S., Erickson, G., & Figueras, N. (2013). International assessments. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 285–302). New Jersey: Wiley and Sons.

Takala, S., Erickson, G., Gustafson, J.-E, & Figueras, N. (2016). Future prospects and challenges in language assessments. In J. Banerjee & D. Tsagari (Eds), *Contemporary Second Language Assessment* (pp. 299–312). London: Bloomsbury.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge: Cambridge University Press.

Turner, C., & Purpura, J. (2015). Learning-oriented assessment in second and foreign language classrooms. In J. Banerjee & D. Tsagari (Eds.), *Handbook of Second Language Assessment* (pp. 255–274). Berlin: De Gruyter

van Moere, A. (2010). Automated spoken language testing: Test construction and scoring model development. In L. Araújo (Ed.), *Computer-based Assessment (CBA) of foreign language speaking skills. Joint Research Centre scientific and technical reports* (pp. 84–99). Brussels: Publications Office of the European Union. Retrieved from: http://publications.jrc.ec.europa.eu/repository/handle/111111111/15037

van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, 29(3), 325–345.

Weir, C. (2005). *Language testing and validation. An evidence-based approach*. London: Palgrave.

Wiliam, D. (2013). *Embedded formative assessment*. Bloomington: Solution Tree.

Received December 17, 2017
Revision received October 31, 2018
Accepted February 5, 2019