Cristiana Cervini, Monica Masperi, Marie Jouannaud & Francesca Scanu

Université Stendhal, Grenoble, France
Università di Bologna, Bologna, Italy

cristiana.cervini@unibo.it

Defining, Modeling and Piloting SELF, a New Formative Assessment Test for Foreign Languages

Bio data

Cristiana Cervini, PHD in educational linguistics, teaches 'applied linguistics to foreign language learning' at the LILEC Dep. (University of Bologna). Her present studies focus on assessment and evaluation and on CALL systems for hybrid and self-learning. She is currently in charge of the SELF research and development actions in the frame of the INNOVA-Langues project (LANSAD, Grenoble 3).

Monica Masperi is a Senior Lecturer in Linguistics and Didactics at Stendhal University, Grenoble, France. At the head of the LANSAD department (Languages for non-specialists) since 2004, she is also the scientific director of the INNOVA-Langues project. Her current research as a member of the LIDILEM research lab focuses on Italian didactics, plurilingualism and the use of technology in language teaching and learning.

Marie-Pierre Jouannaud has been a foreign language instructor for more than twenty years and currently teaches English linguistics and teaching methods at Stendhal University, Grenoble, France. She has a Masters degree in linguistics and is about to start a PhD in applied linguistics. Her areas of interest include acquisition of grammar, blended learning and the development of learner autonomy. She is the English item writer coordinator in SELF.

Francesca Scanu is a teaching assistant in the LANSAD department of Stendhal University - Grenoble 3. She has a Master's degree in Didactics of Italian as a Foreign Language (University for Foreigners, Siena) and is actively engaged in the creation of language learning paths. Within the INNOVA-Langues project, Francesca is currently working on the creation of SELF.

Short paper

Assessing foreign languages in higher education: state of the art and preliminary results

Designing a new foreign language test requires defining what we mean by language and language use (Bachman, 1990). The construct that we use is the cornerstone guiding us when we create original test items and when we design the general architecture of the test. What are the characteristics of the intended test-takers? What do we take 'communicative competence' to be? How can we translate all of this into test items that will be administered online? The Common European Framework of Reference for languages (henceforward CEFR) can guide us but cannot tell us how to anchor our items to the descriptors of each of the skill levels while staying true to task-based and action-oriented approaches to language teaching, and to the role of the learner as a social actor (ALTE, 2011; Weir, 2004).

The purpose of this talk is to describe two aspects of this work in progress:
i) how our team's linguistic, discursive and contextual choices were guided by the principle of situational and interactional authenticity on the one hand, and by the attempt to integrate competences on the other; ii) what the preliminary stages of the piloting process tell us about our students' socio-biographical characteristics and the validity of our test.
SELF (Système d'Evaluation en Langues à visée formative) will eventually cover three language skills (listening, reading and writing[1]), but the first stage of its development focused on listening, because of the high correlation we observed between oral comprehension level and success in foreign language tests. Developing a test to be used in institutional settings implies a series of inevitable constraints due, on the one hand, to the possible wash-back effects on learning and on teaching models and, on the other, to the high number of test-takers taking the test at the same time.

How can we overcome the limits of computer-assisted testing and standardization in foreign language evaluation?
Communicative and task-based/action-oriented approaches require taking into account pragmatic and even sociolinguistic variables, whereas standardized automatic scoring seems more compatible with the testing of discrete linguistic knowledge associated with more traditional methods (morphosyntax, spelling, phonology and lexis).

If we wanted to align closely with the CEFR, we would need to include efficacy and communicative relevance in our analyses, both for monologic (spoken production) and dialogic texts involving two or more speakers (spoken interaction), especially at the higher levels. It is not easy, however, to integrate these fundamental aspects within the constraints of computer-assisted testing: not only will the test have to be automatically corrected (it will be administered to hundreds of students more or less simultaneously during registration week), but it will also need to be relatively short.
Sociolinguistic competences are another source of difficulty. According to the CEFR, being able to identify regional dialects as well as elements of a country's popular culture is one of the skills displayed at higher levels. Unfortunately, considerations of equity prevent us from using dialects and any questions vulnerable to interpretations of cultural stereotyping that might offend some learners or put them at a disadvantage without having anything to do with linguistic competence (Kunnan, 2010).
Some compensatory measures are obviously needed to make sure that our construct of linguistic competence is compatible with the constraints of online testing.

Corpus-based, authentically grounded and home-made items
We define items as 'minimal units of content allowing verification of a linguistic objective'. Most items are self-contained, but their identity is also defined in contrast with or in relation to other items within the system or sub-system they belong to. Of course, the audio document that each item uses determines to a great extent the characteristics of the item. In our case, the three main sources used are:

- Home-made: the item writers create a text (dialogue, news item, …) centered around a communicative, lexical or morphosyntactic element that can thus be specifically targeted because it is hypothesized to be a critical component of linguistic competence at a given level;
- Corpus-based items whose associated audio text came from transcribed corpora of oral language (e.g. the LIP corpus: « Lessico di frequenza dell'italiano parlato »);
- Finally, items whose audio comes from an authentic document whose original purpose had nothing to do with teaching or testing (for example public announcements in a train station).

[1] It might be useful to indicate that exercises should be self-corrective hence we would focus on "limited production" (i.e.: short answers, discourse completion task, etc.).

In order to focus more closely on communicative competence in interaction , one of our item types for listening involves the test taker having to choose the best and most appropriate response for the next turn in an on-going conversation (these items are in a way similar to Dialang's register/appropriacy items in their writing construct and the DCT (discourse completion task) type of exercise).

Situational and interactional authenticity, integration of competences

Language in use does not separate competences, but discrete point testing does. We have tried as much as possible to balance these two contradictory requirements by designing an "identity card" to help us describe, create and classify items. This tool is also essential for the training of new item writers as well as for research and evaluation.
Here are the definitions we use for authenticity and integration of competences:
"Situational authenticity" refers to the accuracy with which tasks and items represent language activities from real life. "Interactional authenticity" refers to the naturalness of the interaction between test taker and task and the mental process which accompany it. [...] To make an item or task more situationally authentic, the key feature of the real life task must be identified and replicated as far as possible".
"Integrating competences": when we are designing a test task, it is important to be clear about the balance between competences needed for a successful response. Some competences will be more important than others - this will form the focus of the task" (Manual, 2011).
At the micro level, enriching each item with exhaustive contextual details aims to make up for the loss of paralinguistic information that is typical of naturally occurring exchanges. More specifically, each item is composed of a series of elements including "contextual clues" given to the test-taker (usually the place where the scene takes place or a short introduction to the topic), the pedagogical direction (explaining what the test taker has to do), which is separate from the functional direction or prompt (the technical operation associated with the item type, such as "choose the correct answer" for a multiple choice item). These indicators complete the input (the audio clip) and the answers (the key and the distractors).

A modular structure that is flexible enough to adapt to different uses

One of the most interesting aspects of the projected SELF system is the variety of purposes for which it is intended. It is designed to be used with non-specialist students (i.e. students who are not majoring in languages) who are taking FL courses to fulfill their foreign language requirement or as an elective. The delivery models for these courses range from face-to-face learning with enriched online content to blended learning or hybrid courses and fully online tutored courses.
SELF will be used as a placement test ("a test administered in order to place students in a group or class at a level appropriate to their degree of knowledge and ability" (ALTE Multilingual Glossary, 1998)) as well as a diagnostic test ("A test which is used for the purpose of discovering a learner's specific strengths or weakness. The result may be used in making decisions on future training, learning or teaching" (ALTE Multilingual Glossary, 1998)). This kind of formative assessment helps students improve their capacity for self-assessment, which, along with greater awareness of their weak and strong points, is the first step towards autonomy. It will also provide information to the tutors who will guide the students during their personalized self-directed online training sessions.
SELF is thus conceived as a modular structure focused on the assessment of three language activities: listening, easing and writing. This modularity will allow for flexible uses. As a placement test, its administration should not exceed 50 minutes or so. As a diagnostic test, it will be possible to assess each skill separately. In this case, its main purpose will be formative, designed to foster learners' autonomy.

What is listening comprehension? items and the cognitive operations they imply

Defining the construct of listening means that we have had to reflect on the stages of the process, the subskills and the cognitive operations involved. Different foci are possible: 'listen and perceive', for bottom-up phonetic and prosodic processing, 'listen and

understand', i.e. (re)constructing and interpreting meaning to grasp the message and 'listen and interact', because efficient and relevant interaction must be based on adequate reception and comprehension of the other speaker's discourse (Cornaire 1998, Nobili 2006).
As far as the difficulty of the task (listening comprehension theory) and the difficulty of items (testing theory) are concerned, the descriptors of the CEFR lead us to identify several contributing factors, namely:

1. the linguistic characteristics of the audio input, i.e. speech rate, pauses, hesitations (phonetic characteristics), or fragments, atypical word order, variation typical of spontaneous oral discourse (morphosyntax), information density linked to the type of text, length and narrative organization of the text (discursive characteristics);
2. the intrinsic difficulty of the exercise type (MCQ, close, T/F, matching, reordering, …), time available to complete the task, use of tools such as the possibility of note-taking;
3. the personal characteristics of the test-taker: ability to make predictions, activation of background knowledge, capacity for sustained attention, verbal working memory, and attitude;
4. the cognitive processes involved, such as listening for gist, listening for details, inferencing about the context (where the scene takes place, who the speakers are, …), recognizing communicative intent (and its effects), identifying mood, register, etc…

The "ID card" of each item, a tool we have perfected as our research progressed, has helped us to:

- Make explicit the focus of the item, so that no ambiguity remains as to what is being tested;
- Raise item-writers' awareness of item characteristics and simplify quality control;
- Facilitate access to specific items with the search engine (by their focus, length, text type, domain, …) and trace their behavior after piloting.
- Make sure there is a variety of focus (lexical, morphosyntactic, communicative), text types and length (long or short, authentic, modified or invented, monologic or dialogic, announcements, instructions, conversations, …), speech rate (fast, medium or slow), language variety (standard or non standard), domains (public, private, professional and educational) and that the test-taker is cast in different roles (listener/eavesdropper or participant).

Two types of tools have turned out to be very useful for the item writers and revisers: reference level descriptions (e.g. 'Il Profilo della Lingua Italiana' for conceiving items in Italian) and oral and written language corpora (such as LIP, CORIS/CODIS, LABLITA).

A tiered approach to the validation of an adaptive test

According to Doucet, the validation process can be likened to an accumulation of converging data until we are convinced that the approach we have chosen is well founded: « Lors du processus de validation, on parlera d'accumulation convergente de données jusqu'à ce que l'on soit convaincu du bien-fondé de l'approche choisie » (Doucet, 2001)
Validatity and reliability are two concepts that come from psychology, where they have been in long use. Reliability implies the stability and coherence of test results over separate administrations (« selon le principe de fiabilité, le résultat d'un test doit rester le même entre deux occasions d'évaluation rapprochées », Lussier & Turner, 1995), whereas validity measures the convergence between the aim of the test and its contents (« le principe de validité sert à vérifier si le test mesure effectivement les performances qu'il cherche à mesurer », ibid.).
We have imagined several steps in the revision and validation process, both qualitative and quantitative, starting with the A2 level listening comprehension items. The qualitative analysis starts with a revision of the items by an expert who is not involved in the writing process, followed by 'think aloud protocols' (TAP) with a few learners selected

because their level in the skill tested corresponds to that of the items studied (this will also enable us to check that the test has face validity for learners).
Scientific validation is then completed by a guided and controlled pilot study where we administer the test to a small group of students of similar ability who serve as a preliminary control group. This is followed by a pretesting phase, in which we observe and interpret statistical data obtained from a much larger pool of learners.

For the think aloud protocol, observing results from two very different groups, one in a second language context (immersion), the other in a foreign language context (non immersion), will allow us to compare their reactions. The first group are volunteer Italian as a foreign language (Lansad) students in Grenoble, France, pretested A2 with Dialang, and the second are Erasmus students in Bologna taking Italian language courses in the Language Center, also pretested A2 using an in-house test. This protocol will help the test designers verify that the thinking processes used correspond to the hypothesized construct, and observe the behavior of the items as well as the impact of our linguistic choices on participants with very heterogeneous biographical and educational backgrounds.
We have focused on the following questions:

- Perceived difficulty of the item (high/ medium/ low), which can be compared with the answer given, whether right or wrong;
- What did the learner find difficult: comprehension of the audio document, of the stem, the answers (key and distractors), the prompt, or was there interference with the native language(s)?
- Is the speech rate of the document too fast, the lexis or structures too complex, the details too hard to follow?
- What is the student's level of confidence in their answer (25/ 50/ 75/ 100%), and would the student have chosen to answer if incorrect answers deducted points out of the total score?
- Comments on the type of protocol and received instructions.

The first results collected during the qualitative analysis have uncovered a bias in one specific multiple-choice item. The oral input consisted in a dialogue between a mother and her son and the question was focused on the interpretation of the mother's feelings. The TAP analysis showed that the mother's behaviour (anxious, angry or thoughtful?) was subject to diverging personal and cultural interpretations, causing a bias. This evidence has led the item-writers to proceed with a second revision and modification of the exercise.

Conclusions: research modules and their perspectives
The SELF evaluation system, a multidimensional tool, will evolve along predetermined lines, the next step being the development of the following modules (or 'research bricks'):

- Self-evaluation module: this step precedes the test proper and has a functional as well as a formative purpose. Functionally, this will allow us to start the test with items closer to the level of the learner, and reduce the time necessary and the stress or boredom associated with excessively hard or easy questions for the test-takers. Formatively, it would be interesting to link the self-evaluation results to the final diagnosis. We are exploring different modalities for self-evaluation, including metacognitive can-do statements ("Je suis capable de…"), benchmarked samples to which the learners can compare their performance, or using ideas developed in portfolio assessment projects, which might mitigate the drawbacks of questionnaires and can-do statements.
- Formative feedback module: feedback is a central issue for a formative test, both for the student and the instructor. This implies deciding what information the student needs to see after finishing the test, and what needs to be stored for the long term, or perhaps permanently, perhaps in a "personal profile" page.

- Scoring module: The protocols we choose for our exercises, the competence levels aimed at, and the use of dichotomous vs. polytomous items will all impact scoring procedures.

References

Bachman L. F. & Palmer A. S. (1996). Language testing in practice: Designing and developing useful language tests, Oxford: Oxford University Press.

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

Cornaire C. (1998). La compréhension orale, CLE International: Paris.

De Mauro T., Mancini F., Vedovelli M. & Voghera M. (1993). Lessico di frequenza dell'italiano parlato, Etas Libri: Milano.

Doucet P. (2001). Pour un test utile. ASp, 34. URL : http://asp.revues.org/1696; DOI : 10.4000/asp.1696.

Kunnan A. (2010). Statistical Analysis for Test Fairness. Revue française de Linguistique appliquée, 16.

Lussier D. & Turner C. E. (1995). Le point sur l'évaluation en didactique des langues, CEC - Centre éducatif et culturel : Montréal.

Manuel pour relier les examens de langues au Cadre européen commun de référence pour les langues. (CECR). (2009). URL http://www.coe.int/t/dg4/linguistic/manuel1_fr.asp.

Manual for Language Test Development and Examining (for use with the CEFR), produced by ALTE (Language Policy). (2011). URL : http://www.coe.int/t/dg4/Linguistic/ManualtLangageTest-Alte2011_EN.pdf.

Nobili P. (2006). Oltre il Libro di Testo. Carocci : Roma.

Rubin J. (1994). A Review of Second Language Listening Comprehension research. The Modern Language Journal. 78, ii.

Weir C. J. (2004). Limitations of the Common European Framework of Reference. Developing Comparable Examinations and Tests, URL: http://ltj.sagepub.com/content/22/3/281.abstract.