18 Designing a Multilingual Large-scale Placement Test with a Formative Perspective: A Case Study at the University of Grenoble Alpes

Cristiana Cervini and Monica Masperi

Abstract An interdisciplinary team composed of more than thirty people has been engaged in the process of designing, developing and validating an online placement test with a formative perspective, called SELF (*Système d'Evaluation en Langues à visée Formative*). SELF is a large-scale assessment system validated according to the ALTE cycle using both quantitative (Classical Test Theory and Item Response Theory) and qualitative methods (questionnaires, interview and focus group), and has been developed within the framework of the ANR IDFI project Innovalangues[1]. Today, SELF has already placed around 120,000 students in six different languages.

Designing a multilingual test with these features is a very demanding and long process. The most challenging aspects concern (1) keeping the same communicative construct for the six different languages; (2) improving item writers' skills in psychometrics and, more broadly, spreading high-quality evaluation culture; (3) infrastructural and technical demands; and (4) coordination of a large, heterogeneous team over a long period of time (six years).

These difficulties required adoption of specific strategies to reach our goal, e.g., careful organization of the working team composed of a scientific manager, team coordinators and item-writers; the decision to start with two pilot languages, Italian and English, followed by the other four; drafting and sharing common documents to guarantee interlinguistic transfer; in-house design of a multi-task platform serving as an authoring tool, a piloting and pre-testing repository, and a large-scale administration system to track, archive and disseminate the final results.

18.1 Introduction: Purpose and Testing Context

The IDEFI-ANR Innovalangues project (Masperi 2011) at the *Université Grenoble Alpes* is concerned with research into innovative pedagogical approaches in the field of teaching and learning second languages. Its main objective is to make a significant contribution to the improvement of language teaching and training practices. One of the central axes of the research is the creation, scientific validation and development of an online formative language assessment system, called SELF (*Système d'Evaluation en Langues à visée Formative*) (Cervini & Jouannaud 2015). SELF is a large-scale assessment system that currently assesses six different languages (English, Italian, Chinese, Japanese, Spanish and French). It

[1] ANR-11-IDFI-0024 – *cf.* 〈hal-02004250〉

C. Cervini
University of Bologna
Bologna, Italy
e-mail: cristiana.cervini@unibo.it

M. Masperi
Université Grenoble Alpes
Grenoble, France
e-mail: monica.masperi@univ-grenoble-alpes.fr

is composed of a set of assessment modules that gauges students' language level based on the Common European Framework of Reference for Languages (CEFR). This specialized system, available to the entire educational community, integrates different functions in the same platform: player, task authoring tool, results manager and test session organizer.

SELF[2] arises from the recognition (and evidence) at national level in higher education in France (Masperi 2011, p. 8) of the inadequacy of operational solutions for formative assessment. The main shortcomings observed include: (1) the closed (non-dynamic) nature of application software; (2) the lack of transparency in the calibration of items used to assess linguistic ability; (3) the absence of tracking of student work; (4) the summary nature of the information provided without any diagnostic assessment that would allow an effective learning response. The evidence of these shortcomings, which are found in all language teaching across the country, encouraged us to propose the ambitious design of a multilingual system to provide guidance and reliably assess the strengths and weaknesses of French-speaking students and so facilitate and provide an incentive for the creation of groups with similar levels and needs (targeted needs-based training).

18.1.1. SELF Conceptual Foundations

The design of an assessment system like SELF must be based on a wide-ranging consideration of the language and skills model to be proposed. In testing, this consideration means defining the construct of the test. In this respect, the CEFR is an important, if not central, point of reference, but insufficient as a guide to designers in the creation of assessment tasks within a communicative approach that respects the level descriptors. From this point of view, the realization of tasks and items must be duly supported by explicit and rigorous procedures that are not set up a priori, but are developed through constant interaction with the academic discipline, the foundations of which have been laid for many years, and a research-action-development approach that operates in a precise area of application.

Specifically, SELF is a teaching tool conceived as a hinge for training that aims to place the student unquestionably at the center of the learning processes. The system is based on the need to adopt the same methodological approach for all target languages in terms of structural coherence, content, assessment processes, and visualization of results. An equally fundamental need is that of realizing a technical and pedagogical system that is both flexible and adaptable, while taking into account the needs of all players involved in learning assessment within institutions, both in teaching (researchers, teachers and students) and in administration.

18.2 Testing Problems Encountered: Communicative Constructs and Standardized Language Tests

The design of SELF is based on a response to a series of key linguistic, pedagogic and organizational questions. The main challenge in the development of the system was to reconcile our pedagogical aims – designing a valid and reliable multilingual communicative test – with the practical constraints linked to standardization and computer-based assessment. A test can be defined as "communicative" if it conveys *meaningful communication exchanges* in *authentic situations* (Brown 2005). Besides these two key points, a real communicative test should have *unpredictable and/or creative language inputs* and *outputs* where *integrated skills* are simultaneously stimulated, as is the case in real life. The features of unpredictability and creativity are the most difficult to reproduce through self-correcting online tasks. An in-

[2] SELF – Système d'Evaluation en Langues à visée Formative

depth definition of the construct and its operationalization can be a valid way to avoid the risk of its under-representation in standardized tests[3].

Before describing the SELF construct in detail, it is important to highlight other relevant constraints in our design. The construct was supposed to be the same for all six languages used, despite different second language acquisition and consolidated testing traditions, which could significantly differ for non-European languages such as Japanese and Chinese (Higashi et al. 2017) compared to Italian, French, Spanish and English, the other four languages included in the system. The first aim of the test system is to guarantee valid and reliable placement in a language course but, given its formative nature, SELF should also provide information and guidance for students and teachers.

Another contextual factor concerns the practicality of the test, which is part of the richer and broader concept of *usefulness* of a test. A test is required to be useful (for institutions, for students, for society in general), and to be useful it should satisfy six requirements: validity, reliability, authenticity, interactivity, impact and practicality (Bachman & Palmer 1996). Practicality within SELF consisted in the design of a durable system for large-scale assessment (more than 120,000 candidates[4] evaluated in around four years), easy and safe to use in an institutional environment. In this specific case, practicality refers not only to the available resources for development and administration (human, economical and organizational), but also – for test candidates – to the reasonable period of time required to complete the test (not more than one hour), considering its low-stake context of exploitation.

SELF's communicative constructs are focused on three abilities – listening, reading and limited production – which means that the principle of interactional or situational authenticity is alternatively based on an oral (just audio, such as a phone call exchange or a radio broadcast, or audio-visual, such as TV news, ads, lessons, etc.) or a written input (e.g., taken from magazines, post-its, newspapers, etc.).

Considering the formative nature of SELF, it is clear that limiting the exploration of language competence to the macro ability does not provide sufficient information for either students or teachers. For this reason, we have expanded some facets of receptive or productive ability through items that we have called "linguistic focalizations" and "cognitive operation(s)." The concept of linguistic focalization partially covers that of a sub-skill, whereas that of "cognitive operation" refers to the process that a candidate is supposed to activate in order to resolve items or to reply to questions. In terms of linguistic focalization, test items concern three main facets of language competence: grammatical knowledge (morphology and syntax), vocabulary (including collocations and idioms) and socio-pragmatic aspects. Due to the multidimensionality of human linguistic expression and of texts, these focalizations often coexist in the same item. In some other specific cases, some items could be more focused on phonetic discrimination, on textuality (coherence and cohesion) or on metalinguistic reflections.

Cognitive operations refer to functions of a subject's cognitive activity, i.e., to the mental processes (understanding, inference) that he/she needs to activate to respond to the item. A cognitive operation also refers to what a candidate is called on to do (complete, interact, correct) with a text that is read/heard. Tracking all these features makes a significant contribution (1) in defining/observing the degree of complexity of the language task and (2)

[3] "Standardized assessment makes a serious effort to capture crucial aspects of the component abilities of comprehension. Drawing on these assumptions for standardized test construction, […] standardized reading assessment should seek to translate (aspects of) the reading construct into an effective reading test (fluency and reading speed; automaticity and rapid word recognition; search processes; vocabulary knowledge; morphological knowledge; syntactic knowledge; text-structure awareness and discourse organization; main-ideas comprehension; recall of relevant details; inferences about text information; strategic-processing abilities; summarization abilities; synthesis skills; evaluation and critical reading" (Grabe 2009, p. 357).

[4] "Around 80% of the administrations were in English, whereas the remaining 20% were more or less equally distributed among Spanish, Japanese, French, Italian and Chinese."

helping to clarify the multidimensional construct of communicative competence. Section 5 will describe the technological measures that we have adopted in order to enhance students' centrality in the testing process and to develop the concept of a formative perspective within SELF.

18.3 Solution of the Problem: The Testing Cycle for a Good Culture in Evaluation

When applied to language testing, the concept of validity has evolved in the last decades in the direction of the study and observation of its social impact on all stakeholders (students, teachers, institutions and society as a whole). Therefore, we have sought to anchor SELF to the best practices in language evaluation, both to afford our system maximum scientific legitimacy and to spread a positive culture in the field of assessment at the University of Grenoble Alpes and within its connected networks. Indeed, validation is not a process to be undertaken on the spur of the moment. It involves a series of different steps which are intertwined and iterative, from quantitative to qualitative and vice-versa. For this reason, it is very important to plan validation well in advance, because the organizational effort required is enormous, particularly in the piloting and pre-testing phases.

These objectives have resulted in some necessary operational choices: (1) invest energy, time and economical resources in acquiring new, specific skills in the field of item writing and psychometrics; (2) increase, through individual and group responsibility and motivation, the team's appreciation of being part of a project with long-term goals to produce a durable system; (3) improve the team's awareness of the risks of subjectivity in language evaluation and, consequently, of its unethical impact on institutions and society.

The main qualitative validation phases at the beginning of the SELF test cycle were (1) content re-reading and peer correction, and (2) think-aloud protocoling to fine-tune the effectiveness of the software interface, whereas at the end of the test cycle, we considered (3) standard setting and post-test qualitative evaluation with teachers and candidates. Generally speaking, "identifying the score which corresponds to achieving a certain level is called standard setting. It inevitably involves subjective judgement, as far as possible based on evidence" (ALTE 2011, p. 44). Different standard setting methodologies exist (focused on learners' corpora, on candidates' performance, on test contents), but for SELF the most adequate was the *bookmark method* (Hsieh 2013), which enabled direct discussion and debate among language teachers regarding features of content (clear and bias-free formulations) and task difficulty based on student competence. The application of the bookmark method for standard setting and post-test analysis encouraged triangulation between intuitive (i.e., assign a level of difficulty to the items during the conception phase), quantitative (large-scale pre-tests and statistical analysis to establish items' psychometric values) and qualitative methods (final validation by experts after reaching a general consensus).

Post-administration analysis was conducted with both students and teachers through questionnaires and interviews. The aim of this analysis was, on the one hand, to discover if students who had been placed in a specific language class on the basis of SELF results felt that they had been placed in the correct group (in terms of proficiency) and, on the other, to assess whether the class group was sufficiently homogeneous, thus making teaching of the class easier for the teachers. In the case of the Italian version of SELF, this qualitative survey proved that the system had a slight tendency to overestimate French students' competence in Italian. This side effect was relatively predictable, because it is a self-corrective test with a strong component based on the evaluation of receptive abilities. This tendency was promptly corrected through two different measures: (1) an increase in the threshold levels for limited production, which was the most discriminating ability in the linguistic combination "Italian for French candidates," and (2) a reduction in the number of reading comprehension items, which proved to be less discriminating than listening and limited production.

Regarding the quantitative methods, it is crucial to remember the fundamental importance of pre-testing both the tasks and the assembled version of the final test. We organized the quantitative validation in two main stages: a *pilot test* on a target corpus of around 50 participants, mainly aimed at improving the quality of content preparation and at providing a first look at item discrimination indexes (we applied the Classical Test Theory through the use of the TiaPlus software), and *pre-testing* on a target group of 250 candidates (this large-scale trial allowed us to apply the Item Response Theory and, in this case, we used the Winstep software). As shown in the ALTE testing cycle, pre-testing was preliminary to item calibration, which, again, occurred before the standard setting phase. Through pre-testing all the items are calibrated and put in order of difficulty but threshold levels have not yet been defined. Therefore, this last step can be accomplished thanks to the new involvement of language teachers or of linguistic experts in the standard setting discussion which is a very interesting process from a cultural and intercultural point of view: teachers and linguistic experts are requested to explicitly uncover and share their vision of language competence with others. Even if competence descriptors are the same for all participants, their interpretation is often very subjective because it reflects individual teaching and learning styles and habits. For this reason, "definition of the threshold scores is probably the part of psychometrics most associated to cultural, political and artistic issues" (Cizek 2011, p. 5).

This very enriching experience revealed the fundamental importance of including direct and indirect users in the test validation cycle, not just to benefit from the natural increase in the social acceptance of the test, but also in order to neutralize bias and other critical issues.

18.3.1 SELF: An In-house Conception of a Multi-task Platform

SELF is a complete system – player, task authoring tool, results manager and test session organizer – that is fully operational and designed to respond to specific needs of research and teaching. Depending on user status (student or editor/administrator), SELF presents two different interfaces. Specifically, the SELF interface enables (1) designers and editors to create tasks and items that they can then assemble into tests, and (2) administrators to manage test sessions and export the results. The different elements that make up a task are shown on the screen always with the same layout and labeling (called the "task grammar"). Shown on the left of the screen (see Fig. 18.1) are (1) the context, (2) the question source (i.e., the input from which the question is formulated), and (3) the number of repetitions (for listening comprehension questions). On the right-hand side, there are: (1) one or more items set out in sequence in tabs, (2) the question, (3) the possible responses, and (4) the button to confirm a response. The combination of all these elements, called "task grammar," has the same features for the three abilities (listening, reading, limited production).

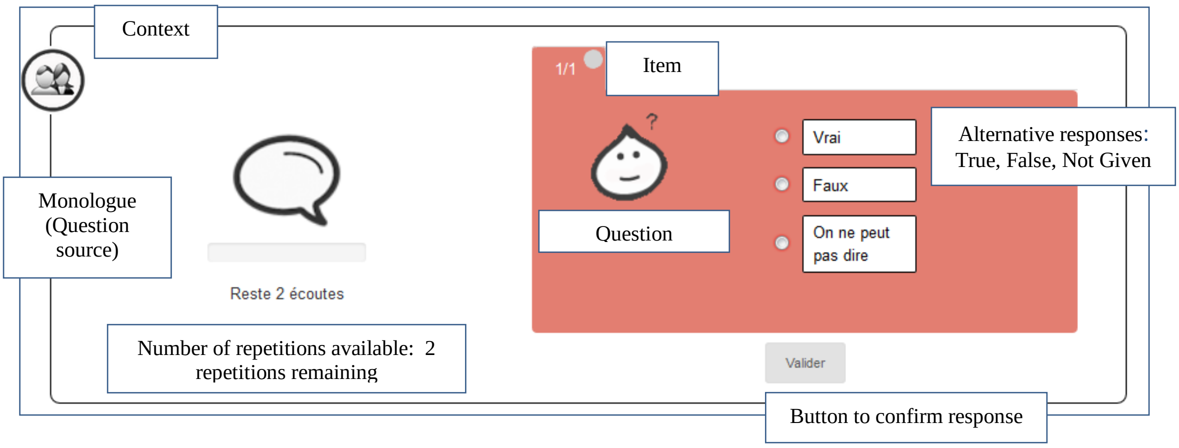
Context
Item
1/1
Vrai
Faux
On ne peut pas dire
Alternative responses: True, False, Not Given
Monologue (Question source)
Question
Reste 2 écoutes
Number of repetitions available: 2 repetitions remaining
Valider
Button to confirm response

Fig. 18.1 Example of SELF layout (for an oral comprehension task)

One of the main strengths of the software is the considerable flexibility in integrating different types of resources. All the fields (context, question source, etc.) are *all media compatible*, so they can accept any audio, visual, image or text source. Regarding task conception, the editor has access to different types of exercise depending on the ability to be tested and the objective of the task[5]. The system flexibility is also linked to the independency of the different item banks. Each test refers to a specific language item bank composed of the validated items, but all six banks (English, Italian, Chinese, Japanese, Spanish and French) are conceived and technically structured in the same way. This feature of the SELF system assures that each language team can easily work in autonomy. In addition, the authoring tool allows editors (1) to integrate *ad hoc* feedback, (2) to retrospectively add to the set of possible responses to a construct "short written expressions" taken from students' responses that are correct but were not initially envisaged by the editor. Finally, tracking of individual and group activity is a powerful added value for researchers and teachers. Even if, at present, only a small part of the information collected with SELF can be exploited to provide feedback to students and teachers, the tracking system used reflects this methodological approach and its relevance for diagnostics and training. The system that tracks and manages results can generate files with different features and objectives. Alongside management and statistical files (the former for administrators involved in organizing groups, and the latter for analysis with statistical software), there are full export files providing a broad range of information relevant to language teaching and learning. This includes some biographic data (e.g., first language(s) and other reference languages in addition to the first language for each candidate), the time required to complete the full test and the time spent on each task before confirming the response, and the level of perceived difficulty (again for each task). Regarding these last two points, correlations in the data collected could give rise to more wide-ranging and highly informative considerations. For example, we could compare the actual difficulty of an item (expressed by the psychometric index of difficulty) with candidates' perceived difficulty and the response (correct or incorrect) given to the question.

[5] "Standardized assessment makes a serious effort to capture crucial aspects of the component abilities of comprehension. Drawing on these assumptions for standardized test construction, […] standardized reading assessment should seek to translate (aspects of) the reading construct into an effective reading test (fluency and reading speed; automaticity and rapid word recognition; search processes; vocabulary knowledge; morphological knowledge; syntactic knowledge; text-structure awareness and discourse organization; main-ideas comprehension; recall of relevant details; inferences about text information; strategic-processing abilities; summarization abilities; synthesis skills; evaluation and critical reading" (Grabe 2009, p. 357).

At the same time, it is important to note that the diagnostic and training approach of SELF is supported by a further tracking tool that we have designed and developed, and entitled *identity card*. The identity card associated with every item and every task tracks essential linguistic and didactic information regarding the characteristics of the written and spoken texts, the specific qualities of individual items, and the psychometric indices. All these factors may influence both task complexity and the strictly individual relationship created between a candidate and the task presented. This is the way in which such a meticulous tracking system can open the door to studies of the diagnostic and training perspective of SELF.

18.4 Insights Gained: Looking back at Process and Choice

The development of a multilingual assessment system founded on a common methodological and didactic framework for use by adult French-speakers was a challenge determined by key, local factors. Today, the widespread dissemination of SELF in academic institutions in France is proof of the need for the tool. However, the results obtained have never been taken for granted. In our opinion, the large-scale adoption of this training tool is based on four joint factors: (1) the quality and stability of the staff involved in the process, (2) the rigorous documentation of the process, (3) the thorough quality control, and (4) the intrinsic nature of the product itself.

First, the work assigned to the designers and editors was of a high professional level (Cervini 2014). From a technical point of view, exceptional linguistic competence must be accompanied by an excellent command of procedures that require a specific training background. However, the process must also include a creative component both in identifying sources and in creating original texts. The role of the *performant item writer* therefore combined a rather uncommon dose of perfectionism and inventiveness. Finally, the collaboration with programmers also had been mediated through a technical and pedagogic professional who defines the profile of the technological, IT and ergonomic specifications.

The second essential aspect was the development and provision of valid and substantial working tools: a reference bibliography, clear and exhaustive methodological and didactic guidelines (regarding text creation, editing of items, and psychometric analyses), interlingual glossaries and didactic memoranda (key words and definitions, types of protocol, question banks), and clearly stated procedures regarding the activities related to the preparation of tasks (studio recordings and use of authentic resources).

In line with the methodology adopted, the third element – quality control of the SELF project – played a role for the six target languages at two process levels: during the creation of the tests and when they were delivered. The measures adopted were of three types: (1) the methodological support for the researcher and editor team provided by international experts in the sector[6]; (2) the product maturation envisaged by the testing cycle and undertaken following the required stages of validation and psychometric analyses; and (3) the compilation of questionnaires during piloting, as well as the ex-post use of qualitative research protocols applied to results collected from the students tested. Moreover, the service offered to the universities using SELF is shown to be appreciated in annual *ad hoc* questionnaires.

Finally, the question of the transferability of innovative teaching practice varies in function of the nature of the product itself. SELF fills an evident gap in the field of learning assessment of which we were fully aware. Moreover, we assume that broad, consensual

[6] In the first stages of development (2013-2015), research methodology was based on suggestions and training provided by CIEP. More recently (2016-2018), the project has enjoyed the expert support of James Purpura (Columbia University) and CITO (Department of Psychometrics and Research), The Netherlands.

adoption of the system might be determined by the fact that SELF is a finished, "turn-key" product that is non-invasive and not in competition with other solutions. SELF might act as a lever to define a university's language policy, but it leaves institutions maximum freedom to decide on the use of their own teaching tools (communicative, action and thematic approaches; classroom, blended and holistic systems, CLIL), and the way in which these tools interact with the assessment system.

18.5 Conclusions: Implications for Test Users

SELF is a multilingual assessment system with a formative and diagnostic aim. It is available in six languages at state higher education institutions in France and is used to quickly assess a student's level in three skill areas. SELF is designed to respond to institutional needs to guide students towards a training path that is suitable to their linguistic profile and thus to facilitate the adequate development of existing expertise. The strengths of the SELF system can be summarized as follows:

Academic Solidity and Interlinguistic Coherence. SELF is based on design procedures and academic validation that are rigorous, from both a quantitative (piloting and psychometric analyses) and qualitative perspective (analysis of references, cross-checked revisions between item writers, standard setting, post-assembly piloting). The task banks are designed and produced following a common methodology for all six languages and aim to guarantee customized teaching. Each task has an "identity card" to categorize both the tasks and the items, indicating the linguistic and pragmatic focus and so serving as a precursor to the diagnostic framework.

Conscientization of Prior Learning and of Perceived Difficulties in a Formative Perspective. In testing each chosen ability, SELF seeks, in spite of the objective limitations of automatic feedback, to propose an assessment context that is as close as possible to an authentic communicative situation. In this respect, we have chosen to assess oral comprehension with a fully oral-based approach without any written support, and to include in the bank of possible responses to written expression questions, any correct responses given by students that were not contemplated by the item editor (Cervini 2016). The formative dimension is further supported by the decision to present results in the form of a "recommended learning path" (e.g., *en route to* …).

Multifunctionality of the Underlying Technical Structure. SELF offers flexible and efficient editing and delivery that can adapt to the needs of different institutions (division of students into groups by academic year, by discipline or by department) and interface easily with training paths set by institutional policy. The system's IT platform, which is currently experimental and could be extensively enhanced, is already able to serve a large number of simultaneous accesses (approximately 500). The technical set up is also designed to serve research (editing tools, analysis and categorization of tasks, archiving of data on user actions and results, tracking of item behavior, etc.) and to respond to developments suggested by the data collected for each language.

The design of a system such as SELF must always be considered as a work in progress and subject to continual improvement, not only as far as the obvious need to update content is concerned, but also regarding verification of the usefulness (validity and reliability) of the test for a body of candidates in continual evolution. Experience has shown that psychometric results from the quantitative assessment can be enriched and complemented with qualitative information collected through interviews, focus groups and questionnaires.

The methodological framework that has been established for the development of SELF's diagnostic and formative perspective is specifically based on modelling this information to the benefit of language students (self-awareness, motivation, customized learning paths) and language teaching staff – teachers and tutors – who can more easily design remedial work appropriate for student needs.

References

ALTE (2011). *Manuel pour l'élaboration et la passation de tests et d'examens de langue.* Division des Politiques Linguistiques. Strasbourg: Conseil de l'Europe, DG II – Service de l'éducation.

Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brown, J.D. (2005). *Testing in language programs*. New York: McGraw-Hill.

Cervini, C. (2014). La valutazione multilingue nel contesto dei dispositivi formativi: Il sistema 'SELF' per il posizionamento e la diagnosi delle competenze linguistiche, *LEND – Lingua e Nuova Didattica. Periodico di Linguistica Applicata e Glottodidattica,* 1, 16-26.

Cervini, C. (2016). Approcci integrati nel testing linguistico: Esperienze di progettazione e validazione in prospettiva interlinguistica. In C. Cervini (Ed.) *Interdisciplinarità e apprendimento linguistico nei nuovi contesti formativi. L'apprendente di lingue tra tradizione e innovazione*. Bologna: Quaderni del CESLIC. http://amsacta.unibo.it/5069/1/Volume%2520CeSLiC.pdf. Accessed 18 February 2019.

Cervini, C., & Jouannaud, M.P. (2015). Ouvertures et tensions liées à la conception d'un système d'évaluation numérique multilingue en ligne dans une perspective communicative et actionnelle. *ALSIC – Apprentissage des langues et systèmes d'information et de communication. Numéro spécial 'Des machines et des langues', Alsic,* 18, 2. http://alsic.revues.org/2821. Accessed 18 February 2019.

Cizek, G.J. (Ed.) (2011). *Setting performance standards: Foundations, methods, and innovations.* New York: Routledge.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Higashi, T., Shirota, C., Nagata, M. (2017). Developing a Japanese language test for a multilingual online assessment system: Towards an action-oriented approach to Japanese instruction in Europe. Bologna: *Alte 6Th International Conference Proceedings* (pp. 236-245).

Hsieh, M. (2013). Comparing yes/no Angoff and bookmark standard setting methods in the context of English assessment. *Language Assessment Quarterly, 10*(3), 331-350.

Masperi, M. (2011). Innovalangues: Innovation et transformation des pratiques de l'enseignement-apprentissage des langues dans l'enseignement supérieur. MESRI. ANR. Investissements d'avenir. https://hal.archives-ouvertes.fr/hal-02004250. Accessed 18 February 2019.