Quality Assurance and Foreign Languages - Reflecting on oral assessment practices in two University Spanish Language Programs in Australia

Adriana R. Díaz
University of Queensland, a.diaz@uq.edu.au

Hugo Hortiguera
Griffith University, H.Hortiguera@griffith.edu.au

Marcia Espinoza Vera
University of Queensland, m.espinozavera@uq.edu.au

Quality Assurance and Foreign Languages - Reflecting on oral assessment practices in two University Spanish Language Programs in Australia

Abstract

In the era of quality assurance (QA), close scrutiny of assessment practices has been intensified worldwide across the board. However, in the Australian context, trends in QA efforts have not reached the field of modern/foreign languages. This has largely resulted in leaving the establishment of language proficiency benchmarking up to individual institutions and programs of study. This paper discusses the findings of a cross-institutional collaborative research project focused on the comparative analysis and review of assessment practices in the Spanish language majors at the University of Queensland (UQ) and Griffith University (GU), both members of the Brisbane Universities Languages Alliance (BULA). The project had a two-pronged focus; on the one hand, establishing comparable student academic achievement standards, specifically for oral assessment in intermediate level courses; and, on the other hand, providing tools and resources to train teachers (continuing and sessional staff) in consensus moderation (CM) practices through an online platform. The results presented here offer practical pedagogical suggestions to support planning and review of oral assessment, thus contributing to QA management in languages other than English.

Keywords

Quality Assurance and Foreign Languages - Reflecting on oral assessment practices in two University Spanish Language Programs in Australia

Adriana R. Díaz
University of Queensland, a.diaz@uq.edu.au

Hugo Hortiguera
Griffith University, H.Hortiguera@griffith.edu.au

Marcia Espinoza Vera
University of Queensland, m.espinozavera@uq.edu.au

Quality Assurance and Foreign Languages - Reflecting on oral assessment practices in two University Spanish Language Programs in Australia

Abstract

In the era of quality assurance (QA), close scrutiny of assessment practices has been intensified worldwide across the board. However, in the Australian context, trends in QA efforts have not reached the field of modern/foreign languages. This has largely resulted in leaving the establishment of language proficiency benchmarking up to individual institutions and programs of study. This paper discusses the findings of a cross-institutional collaborative research project focused on the comparative analysis and review of assessment practices in the Spanish language majors at the University of Queensland (UQ) and Griffith University (GU), both members of the Brisbane Universities Languages Alliance (BULA). The project had a two-pronged focus; on the one hand, establishing comparable student academic achievement standards, specifically for oral assessment in intermediate level courses; and, on the other hand, providing tools and resources to train teachers (continuing and sessional staff) in consensus moderation (CM) practices through an online platform. The results presented here offer practical pedagogical suggestions to support planning and review of oral assessment, thus contributing to QA management in languages other than English.

Keywords

Introduction

The international higher-education landscape is currently characterised by the need to establish accountability and, in so doing, to explicitly rationalise learning outcomes and academic-achievement standards to support quality assurance (QA). The preoccupation with QA that currently drives policy is only set to intensify in the future (Egron-Polak & Hudson 2010, p.137). However, as noted by Díaz (2013), trends in QA efforts have not reached the language-studies field, or at least they have done so with inconsistencies at the international, national and institutional levels. In some cases, this has largely resulted in leaving the establishment of benchmarking up to professional associations. This means that the overall articulation and regulation of student academic-achievement standards in modern/foreign language programs in higher education is still at a rudimentary stage (cf. Heyworth 2013 for a current review of the literature on quality management and language education).

Since 2011, in Australia, the Tertiary Education Quality and Standards Agency (TEQSA) has been the national regulatory and quality agency that promotes, audits and reports on quality assurance in Australian higher education. This regulatory role was formerly the responsibility of the Australian Universities Quality Agency (AUQA). Currently, the only available language-related guidelines provided by AUQA and TEQSA focus on the development of English-language education for international students. The AUQA report "Good Practice Principles for English Language Proficiency for International Students in Australian Universities", funded by the Department of Education, Employment and Workplace Relations (DEEWR), acknowledged that "there is also an increased recognition within universities of the fundamental nature of language in learning and academic achievement for all students" (AUQA 2009, p.2). Yet efforts continue to focus on areas of direct economic relevance to universities; that is, the satisfaction of full-fee-paying overseas students (Harris 2013). While ensuring the development of English-language proficiency for international students is certainly an imperative, giving it such priority at the expense of other areas of language teaching is a clear indication of the often narrow, one-dimensional conceptualisation of QA processes for languages held not only by many institutions, but also by the national government (Heyworth 2013). This type of one-dimensional perspective seems to be supported and, at the same time, to perpetuate the monolingual mindset embedded in Australian society (Clyne 2005; Nettelbeck et al. 2007; Clyne et al. 2007).

Higher-education institutions around the world are expected to have a structured system to develop, monitor and validate existing programs, but the lack of attention these programs receive largely results in relaxed accountability for the sector. This does not mean that individual institutions do not apply internal QA measures such as reviews and evaluations of programs and subjects (Heyworth 2013). But it does suggest that: 1) there is currently no way to compare or externally validate outcomes related to language learning across countries (except perhaps within supranational levels such as the European Union), and certainly not at the national level in Australia; and 2) it is up to individual language faculties, departments and, ultimately, language teachers themselves to monitor intercultural graduate outcomes. This paper discusses our attempt to tackle the latter in two Spanish-language programs at two different higher-education institutions.

The project discussed here presents a snapshot of the ongoing work being implemented within (cf. Fenton-Smith & Walkinshaw 2014) and across institutions. This paper focuses on the collaboratively underpinned comparative analysis and review of assessment practices in the Spanish-language majors at the University of Queensland (UQ) and Griffith University (GU), both

members of the Brisbane Universities Languages Alliance (BULA) in Australia. The purpose of this project was twofold. It intended to establish comparable student academic-achievement standards, specifically for oral assessment in intermediate courses. It also aimed to provide tools and resources to train teachers (continuing and sessional staff) in consensus moderation (CM) through an online platform. In other words, it focused on both benchmarking and identifying instances of good practice in oral assessment for continuing professional development, two key QA exercises in languages education (Boiron & Mureşan, 2007).

Collaborative Project Background and Objectives

The Brisbane Universities Languages Alliance (BULA) is "the largest current collaborative venture" to deliver languages across institutions in Australia (UQ, GU and Queensland University of Technology), supported and encouraged by the Australian government's Collaboration and Structural Reform (CASR) Fund (2004-2007) (Dunne & Pavlyshyn 2012, p.14). The Spanish programs at UQ and GU, both of which were established approximately 20 years ago, share several features. Both use the same set textbook, *Dos Mundos* (Terrell et al. 2010); the same sequence of content across the first two years of study, elementary and intermediate levels; similar delivery modes (an average of four contact hours per week); and the same teaching methodology (a combination of natural and communicative language-teaching approaches). This made them ideal candidates for the type of collaborative research encouraged through BULA; specifically, the *BULA Scholarship of Teaching and Learning Grant Scheme* provided the ideal springboard to consider a collaborative approach to reviewing assessment practices internally (within programs) and externally (across institutions). Our project was directly aligned with one of BULA's main objectives: to enhance and expand collaboration across institutions with a view to promoting and sharing the application and development of best pedagogical practices (Levy & Steel 2012). In so doing, it also addressed two QA goals that were key priorities for both institutions: enhancing student-centred practices and improving transparency through the external moderation of coursework assessment.

Like many institutions around Australia (cf. Normand-Marconnet & Lo Bianco 2013), both institutions currently map their courses' language-proficiency benchmarking against the Common European Framework of Reference (CEFR). This framework includes a scale of six levels that is recognised worldwide to compare achievements and learning across languages, from A1-A2 (Basic User), through B1-B2 (Independent User) to C1-C2 (Proficient User). Nevertheless, a close look at the oral and written assessment instruments and practices in both universities revealed a number of differences, particularly in the quantity and frequency of the assessment tasks. The UQ Spanish program emphasises continuous assessment, particularly of written tasks, with only one oral exam at the end of the semester. The GU Spanish program, on the other hand, includes an oral exam and a written exam half way through the semester, and again at the end of the semester.

The Spanish programs at both institutions share the understanding that the assessment of speaking skills has significant implications for students and teachers alike. Indeed, it is sometimes claimed that testing second-language speaking is much more difficult than testing other second-language skills, "perhaps because is the ability that makes us human. Perhaps because speaking is fleeting, temporal and ephemeral" (Fulcher 2003, p.xv). As such, the importance of designing speaking tests to be relevant and meaningful for learners is widely acknowledged: "[t]he constructs should be driven by test purpose, taking into account the desires and motivations of those who will take the test, and be sensitive to the requirements of score users" (p.23). In our case, assessment tasks

included oral presentations and interviews, with role-play discussion of graphic material given to the students prior to the test and of topics related to the material taught in class.

While both programs share relatively similar emphasis on oral skills and testing instruments, a closer look at the *administration* of oral-assessment activities revealed markedly different descriptors and marking rubrics (holistic and analytic), as well as different internal consensus-moderation strategies. While both programs conduct routine "double-marking", the main strategy to ensure the reliability of oral tests is "sample marking" at GU and the use of two examiners (interviewer and assessor) at UQ. Both institutions acknowledge that moderation processes are required in order to achieve reliability in teacher judgement; in this context, consensus moderation – entailing teachers getting together and meeting to compare, exchange and discuss students' work samples – is one of the most common practices. Although this system is time-consuming, it is an effective way for teachers to gain experience in, and consider the ongoing adjustment of, their evaluative knowledge with respect to student oral performance. Another aspect to be considered was the implementation of consensus-moderation practices within and across institutions, as well as the benefits in terms of teachers' reflective practices and their continuing professional development.

In this project we thus followed the stated objectives and principles underpinning the LanQua Quality Model, funded with support from the European Commission, to reflect on teaching practice and enhance the quality of the experience in the learning and teaching of languages in higher-education institutions. As the LanQua Model suggests, QA should be developed *by* higher-education teachers *for* higher-education teachers "to stimulate reflection and discussion, and provide tools which can be used to demonstrate good practice to a range of other practitioners" (LanQua 2010, p.2).

At an overarching level, the project we conducted was driven by a student-centred approach to learning and teaching, and ultimately aimed to facilitate students' successful achievement of leaning outcomes. To that end, it also focused on teachers' professional learning and their development as calibrated markers. Indeed, by providing teachers (continuing and sessional) and learners with a shared understanding of academic-achievement standards that would be comparable across institutions, we aimed to enhance the overall quality of teaching, the level of student satisfaction and, as a result, student engagement and retention rates. Moreover, the creation of an online platform to share best practices and consensus-moderation tools and resources enabled us to train current and new sessional staff joining both teaching teams.

The main project objectives were classified under three main categories:

1) Objectives for cross-institutional collaboration:
   a. to enhance and expand collaboration with a view to promoting and sharing the application and development of best pedagogical practices (Levy & Steel 2012);
   b. to comply with quality assurance strategies focused on enhancing student-centred practices and improving transparency through the external moderation of coursework assessment, as per TEQSA's (2011) preliminary statement on teaching and learning standards.

2) Objectives for the Spanish-language programs:
   a. to examine and review assessment tasks (in particular, oral exams), and develop comparable learning outcomes and academic-achievement standards between relevant cognate courses within and across institutions;

     b.   to establish general agreement about the quality of these outcomes and standards to ensure that the judgements of students' performance are consistent and have the same "meaning".

  3) Objectives for the participating language teachers:
     a.   to reflect on their current beliefs about assessment and consensus-moderation practices, particularly in relation to oral tasks; and
     b.   to gain a better understanding and knowledge of consensus-moderation practices to become calibrated markers.

Methodological Approach and Data Analysis

The methodological framing of this project was broadly embedded in the Action Inquiry paradigm. As Tripp (2003) explains, this paradigm is an umbrella term for the deliberate use of any kind of a plan, act, describe, review cycle for inquiry into action in a field of practice. Reflective practice, diagnostic practice, action learning, action research and researched action are all kinds of action inquiry.

The project's methodology was underpinned by the Action Inquiry-inspired PIRI Model (Plan, Implement, Review and Improve), as it is a well-established, continuous QA strategy already in use at GU. The planning and implementation phases of the project resulted in revisions to the overall number and type of assessment tasks in each of the programs, as well as the development of revised marking tools (rubrics and descriptors) for the assessment of oral exams.

In keeping with this model, the Review stage included the following data sources suggested by Smith's 4Q Model of Evaluation (2008):
- Teacher self-reflection (teacher reflective questionnaires before and after the consensus-moderation workshop).
- Teacher peer-review (team members and tutors' feedback and focus-group discussion).
- Student learning (student results).
- Student experience expressed through institutional course-evaluation tools – Student Evaluation of Course and Student Evaluation of Course and Teacher.

The project was divided into four main phases:
1. Analysis of the state of play
2. Working towards grade integrity
3. Development of training tools and resources
4. Implementation and Evaluation

The following paragraphs present the main activities relevant to each phase as well as a discussion of the main corresponding outcomes.

Phase 1 – Analysis of the state of play

The first phase entailed a comparative analysis regarding current assessment practices in elementary and intermediate levels of the Spanish-language majors at both universities. This was conducted as follows:
1. A peer review of the alignment between relevant courses' learning objectives and their relation to our curricular content. This entailed an attentive revision to our learning targets

defined in our course profiles and a reflection on our expectations and factual results. Based on our analysis, we identified possible inconsistencies and made corresponding adjustments to our sequence of learning objectives, content and assessment practices in relevant courses.

2. A discussion and review of the assessment tasks, marking criteria and academic-achievement standards used to judge our students' achievements, with particular attention to our oral-assessment tasks.

3. A discussion and review of our students' performance at elementary- and intermediate-level courses. Samples of oral exams from both institutions were examined to identify comparable learning outcomes and academic achievement standards between relevant cognate courses within and across institutions.

4. A detailed comparison of the performance criteria for oral exams in Spanish used by institutions from Australia and the United States (where the teaching of Spanish language to English-speaking background students has a long-established tradition). We analysed the assessment descriptors from the following institutions:

- Griffith University
- University of Queensland
- NSW Board of Studies
- Massachusetts Tests for Educator
- New York University
- North Carolina State University

The parameters analysed were:
- learning objectives;
- dimensions or traits of performance;
- defined linguistic structures expected from students;
- gradations of the criteria and indication of the worth of the performance; and
- common problems observed.

Phase 2 – Working towards grade integrity

In this phase we focused on tasks aimed at ensuring that *grade integrity* principles (cf. Sadler 2009; Sadler 2013) were upheld in Spanish-language oral-assessment tasks. This phase entailed:

Comparability: Assessment tasks, criteria, processes and outcomes from both programs were thoroughly examined to determine if they had enough common ground, equivalence, or similarities to permit a meaningful comparative analysis. Areas analysed were: 1) course design and content; 2) contact hours; 3) learning outcomes; 4) individual assessment tasks; 5) scoring criteria; 6) rating scales; and 7) holistic versus analytical scores.

Commensurability. We analysed the assessment criteria for our oral exams to justify the award of the given marks. We then discussed the correspondence between the marks awarded, the general statements and dimensions formulated to define the qualities for our marks. Preliminary results of our analysis of assessment criteria showed some discrepancies between programs. Therefore, we developed a series of strategies for improvement: 1) a reflection on the correspondence between the mark awarded and the level of achievement within approved cut-off marks; 2) joint planning to decide the parameters of new rubrics to be used in both programs; 3) collegial discussions to distinguish effective from ineffective oral responses; and 4) formative evaluation.

Consensus moderation. We selected elements suggested by Sadler (2009, p.821) to crystallise and convey academic-achievement standards: exemplars, explanations, conversations and acknowledgement of teachers' tacit knowledge on their judgements. We gathered oral performance samples (oral interview tests) to determine the characteristics of the students' oral production and distinguish the dimensions of the rating scales. We comprehensively examined these samples to ensure that the judgements of students' performance were consistent and had the same "meaning" irrespective of time, place or marker. Exemplars, explanations, conversations and acknowledgement of teachers' tacit knowledge on their judgements were also discussed. Based on the feedback and discussions, we prepared a first draft of the descriptors at each point of the rating scale, containing gradations of the criteria and some indication of the worth of the performance, including evaluative components. This process replicated several aspects underpinning the consensus-moderation practices established in English Language Enhancement Courses (ELEC) courses within the School of Language and Linguistics (LAL) (Michael 2011).

Phase 3 – Development of training tools and resources

In addition to the revised assessment tasks and set of academic-achievement standard descriptors, the main deliverable of this project was the creation of an online platform containing samples and exemplars of student oral-assessment production as well as ready-to-use banks of practice assessment tasks and corresponding moderation and marking guides for both immediate and developmental purposes.

After considerable research, we found a suitable platform: [CourseSites]. This is a free, interactive, web-based course-creation and facilitation service that allows university instructors to create and update course material while promoting collaboration and interaction. This platform is powered by Blackboard's current technology (including Blackboard Learn™, Blackboard Collaborate™, Blackboard Mobile™, and Blackboard Connect™). As it was a Blackboard product, staff at both universities were already familiar with its overall layout and functionality. The platform allows the enrolment of participants from any institution (simply by using an email address). This enrolment system also allowed us to choose between different levels of accessibility ("students" or "instructors") to a password-protected environment for all the materials and resources we selected. In addition, like Blackboard, it supports a variety of languages, including Spanish, which we were able to use for our course: *BULA Consensus Moderation Training*.

The course site was divided into four main sections (Figure 1).
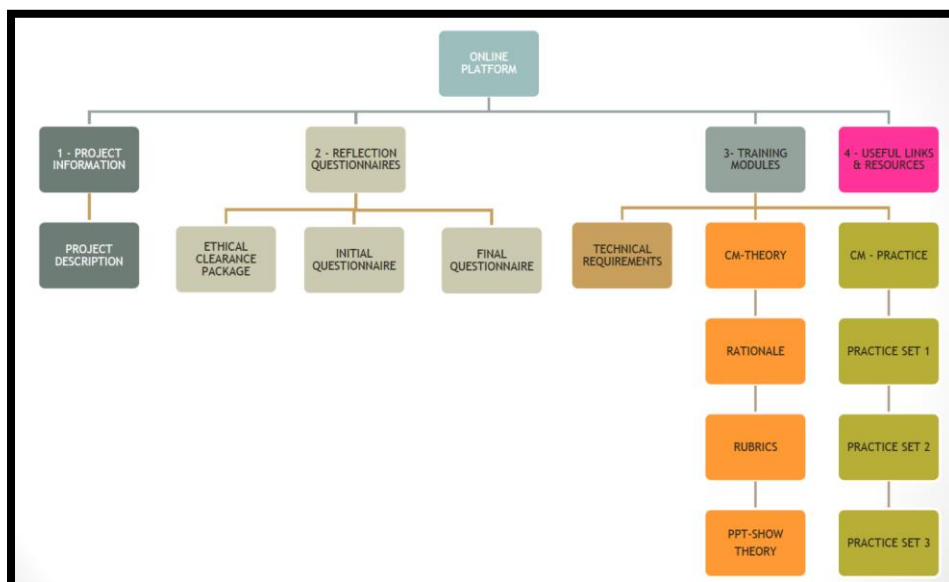
Figure 1. Visual representation of the online platform sitemap

- Project Information –The main aims and objectives of our research project. (Figure 2 shows a screen shot of this page).

- Reflection Questionnaires – Two online anonymous questionnaires, one before and one after completion of the training modules and face-to-face session. This section also provided participating teachers with the corresponding project's Ethical Clearance Information and Consent Package.

- Training Modules – Two main training modules (one theoretical and one practical). The theoretical module included the newly developed set of academic-achievement standard descriptors and holistic and analytic marking rubrics as well as a video explaining their rationale and functioning. The practical module included three "Practice Sets", or exemplars of student oral-assessment production, with their corresponding analysis.

- Useful Links & Resources – A number of key resources in assessment design and best practices in consensus moderation.
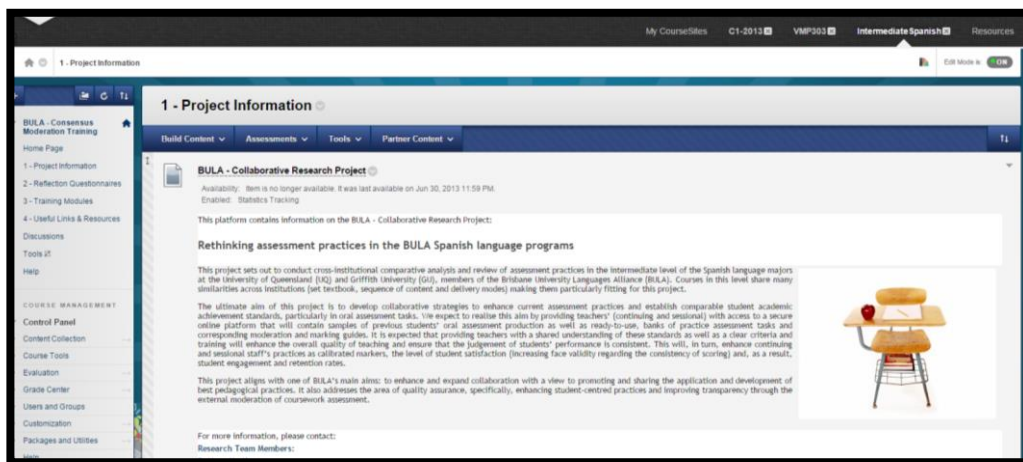
Figure 2. Screenshot of the project's information page

We conducted a consensus-moderation training session with sessional and full-time colleagues (four sessional teachers, including two from each institution, and three full-time colleagues from UQ). The aims of this workshop were: 1) to present a newly developed set of academic-achievement standard descriptors for the final oral assessment in Intermediate Spanish, Semester I courses at both institutions; and 2) to conduct a consensus-moderation session based on these descriptors.

The workshop was divided into two main components: an online component to be completed at home (approximately four hours) and a face-to-face component (three hours). The online component entailed a) accessing an online platform containing relevant training material (including a brief video explanation of the new set of academic-achievement standard descriptors) and b) completing practical exercises to ensure familiarisation with these descriptors (including trial-marking of sample student work).

Before attending the face-to-face session, we gave the teachers a new set of oral exams to assess. At this meeting we discussed and reach a shared understanding, or consensus, about the analytic and holistic marks assigned to each student's performance. The teachers also completed two reflective online questionnaires, one before and one after the completion of the workshop. The questions were related to their perceptions of current assessment practices in the courses they have been involved in as well as their experience in the consensus-moderation process. The estimated completion time of these questionnaires was approximately 30 minutes each (one hour in total).

We will continue to use this online platform and its tools and resources to train current and new sessional staff joining the teaching teams so that they can comply with assessment procedures.

Phase 4 – Implementation and evaluation

Regarding impact on student learning – in other words, students' results – we were particularly interested in determining whether changing the nature and number of assessment items had an

impact on the average distribution of final grades. We expect this to be an ongoing process, as it warrants detailed, longitudinal statistical analysis to yield data suitable for effective comparison.

Students' experience data provided information about the overall assessment design. In this regard, students' level of satisfaction seemed to remain stable compared to previous semesters. At GU, this was reflected responses to Question 2 of our Student Evaluation of Courses, *"The assessment was clear and fair"*, which received a mean average of 4.4 out of 5. Question 3, "*I received helpful feedback on my assessment work*", revealed an increase from an average of 4.1 to 4.4. We intend to continue exploring students' experience data in future offerings through complementary qualitative data-collection tools (e.g., students' focus-group discussion on assessments in relevant courses). As with student results, these figures require ongoing, long-term examination to yield data suitable for effective comparison.

Data collected from teacher-participants, teacher self-reflection (teacher reflective questionnaires before and after the consensus-moderation workshop)[1] and peer review (team members' and tutors' feedback and focus-group discussion) provided several insights into the overall consensus-moderation process. The data yielded by the questionnaires indicated that the consensus-moderation online platform and the face-to-face session were well-received by teachers. In particular, teachers, whose comments are anonymously identified here with a number (T1= Teacher 1, etc.), appreciated the opportunity to reflect on the complexity of oral assessment and the application of newly developed marking tools:

> The assessment of audio files was the most interesting activity (and was, as you can imagine, the most complex). It is the activity that not only integrates all previous concepts explained, it stimulates reflection on the process of evaluation, especially in regards to borderline cases, which require refined judgment and more subtle use of the criteria to find specific differences.(T1)

During the face-to-face session teacher-participants indicated that they found it particularly beneficial to share their views and justify their judgements with colleagues across institutions. They were appreciative of the group reflection process which provided them with the opportunity to learn about each other's ideas and strategies on how to deal with the challenges of developing and implementing marking tools.

> I think it was a very productive process to both question/reflect on my own assessment practice and have the opportunity to hear other colleagues' opinions during the workshop. (T3)

> What I liked was the exchange with colleagues, which allowed me to consider other points of view and also confirm aspects that I agree with the most. (T4)

To improve future iterations of this session, participating teachers mentioned the need for additional practice samples to illustrate a wider range of cases, including more outliers and borderline cases; such samples should include a variety of examiners, possibly from both institutions (our current oral-test sample pool only includes tests from a UQ examiner). In addition, they also suggested having an additional meeting to repeat the consensus-moderation cycle.

---

[1] When answering the reflective questionnaires, participating teachers had the option to answer in either English or Spanish. For the purpose of this paper, all comments in Spanish were translated into English by the researchers.

Overall, participating teachers highlighted the following aspects of the consensus-moderation process:

> For me, participating in this CM [consensus-moderation] process has been like a refresher course bringing renewed light on oral assessment by studying descriptors and having the opportunity to spend time analysing more detailed assessment criteria. I have always considered assessing oral skills a very difficult matter, but after this CM process I have clearer criteria on which to base future assessment. (T1)

> I think the development of this project achieved significant results...it led to reflective, individual work, subsequently enabling a common ground for comparing and agreeing on the application of the evaluation criteria. Overall, the project contains innovative aspects in relation to the reconceptualisation of assessment categories [for] oral activities. (T3)

Finally, participating teachers' responses also pointed to the ongoing nature of developing and implementing marking tools and criteria in cognate courses. This is illustrated in the following comment:

> It was a great experience, very well organised, both in terms of the materials prior to the session as well as the pool of students' work. It was certainly very useful for me to see similarities and differences in the evaluation process and test the suitability of the different criteria in the marking template provided. The experience has helped me to implement an adapted version of this template in my own intermediate Spanish level course, largely increasing the reliability of the assessment task. (T5)

With regard to the development and implementation of marking tools, this project enabled us, along with the participating teachers, to revise current marking methods and tools; this process, however, presented several challenges. While the holistic matrix describing the academic-achievement standards was created with a relatively high degree of consensus within the research team, the analytic-marking matrix presented difficulties, not least the different weight (importance) attributed to each of the selected dimensions (fluency, accuracy, etc.) and the rating scales (0-100 versus 0-5). In this context, the dimension of pronunciation was particularly problematic, with contrasting positions regarding its weight across institutions, a reflection of current debate in the area (Steed & Delicado Cantero 2014), which is unlikely to be resolved without ongoing research.

Despite these differences, Spanish programs at both institutions are committed to continuing work on adapted versions of these holistic and analytic rubrics. We expect, for instance, that ongoing implementation cycles will enable us to find increasingly succinct descriptors for the holistic matrix and gradually increase the reliability of the analytic rubrics. In addition, we also intend to review our assessment-feedback strategies to match the revised criteria. All these aspects remain to be further investigated in each of the institutions' educational contexts, teaching teams and various levels of proficiency.

As a result of this assessment-review process, changes were made at the macro (programmatic) level and the micro (course) level. While maintaining the CEFR as a guideline against which to map the development of language-proficiency benchmarking, programs at both institutions

changed the number and type of assessment items in elementary- and intermediate-level courses. At GU, a 10% quiz was added in week 4. The addition of this assessment item aimed to provide an opportunity for teachers and students to gauge their learning progress earlier in the semester. This is consistent with institution-wide assessment design enhancement strategies (Wilson & Lizzio 2011). At UQ, a mid-term oral exam was added in week 8 of the semester. This new exam consists of a "role play" in which students are required ask each other questions related to the content studied during the first seven weeks of the semester. This type of peer-to-peer assessment has recently been linked to claims of positive follow-on effects in the classroom, as it is considered to be more representative of classroom interactions (Ducasse 2010; Ducasse & Brown 2009). Additionally, from a practical perspective, this type of assessment is generally considered to be "more time and cost efficient as candidates are tested together, and raters assess two or more candidates simultaneously" (2009, p. 424).

Conclusion

Based on the outcomes and findings discussed above, all three sets of objectives set out for this project were successfully met, particularly in terms of teachers' critical reflective practices and professional learning. Overall, this project enabled us to:

- externally evaluate the effectiveness of assessment practices and marking tools;
- formulate suitable academic standard descriptors and establish mechanisms for their continuing improvement;
- ensure ongoing professional learning through projected cycles of inquiry and maintenance of and improvements to the current online platform;
- improve the quality of the Spanish curriculum by reviewing its relevance to best reflect current professional practice;
- enhance the overall quality of teaching and the level of student satisfaction by facilitating students' successful achievement of learning outcomes and providing feedback that is clear, informative, timely and relevant; in the long term, we expect that this will enhance student engagement and retention rates.

Nevertheless, we acknowledge that ongoing research is required to ensure the sustainability of these outcomes. Indeed, ongoing cross-institutional collaboration within and across state lines in Australia would also contribute to strengthening QA measures in languages education. For instance, initiatives such as the LanQua project in Europe (LanQua 2010) have the potential to support greater sharing of practice in QA across languages and institutional contexts. Macro-level QA initiatives like this one would also have to work congruently with micro-level, teacher-driven strategies like the one presented in this project. Indeed, a critical dimension of QA has to do with teachers' accountability in relation to the continuous improvement of their practice. In this context, institutional support for teachers' classroom-based learning is indispensable to the sustained provision of high-quality education. In addition, as discussed above, institutional support should also be provided in terms of professional-development opportunities that may provide a framework for the type of bottom-up, teacher-driven curriculum reform required for the sustainable innovation that lies at the heart of QA.

Additional research also remains to be conducted into the mechanics of teachers' critical reflections on their assessment practices and the development and implementation of marking tools. Indeed, these areas are two aspects that have been highlighted in recent literature as requiring additional research. Future iterations of the consensus-moderation process would thus allow for such research to be conducted. As such, we expect that the work presented here will

provide a basis for ongoing review of assessment-design practices and development and implementation of descriptors and rubrics in intermediate level courses, and for expansion of this review to the elementary and advanced levels. In addition, while this cycle of inquiry focused on the development and implementation of marking tools, future cycles will also explore the development of assessment-feedback strategies. Finally, we expect that additional cycles of inquiry currently underway will also enable us to bridge the gap between beginning and intermediate levels of instruction and assessment tasks, both oral and written.

On a personal level, our own perspective on assessment-design practices has been transformed through this investigation. In helping participating teachers examine their own beliefs and judgements, we have critically reflected on our own professional learning and the type of pedagogical competences needed to continue enhancing assessment practices in our programs of study. We hope this project and subsequent cycles of inquiry will contribute to the burgeoning development of the "scholarship of assessment" field in higher education (Rust 2007; Banta 2002), particularly in the largely unexplored context of languages other than English, where bridges to the well-established field of language testing should be built (cf. McNamara 2014), all framed within current QA imperatives.

References

AUQA (2009). *Good Practice Principles for English Language Proficiency for International Students in Australian Universities*, Department of Education, Employment and Workplace Relations, Canberra.

Banta, T W (2002). *Building a Scholarship of Assessment,* San Francisco: Jossey-Bass.

Boiron, M & Mureşan, L (2007). *QualiTraining: A Training Guide for Quality Assurance in Language Education*, Council of Europe Publishing, Strasbourg.

Clyne, M (2005). *Australia's Language Potential,* UNSW Press, Sydney.

Clyne, M, Pauwels, A & Sussex, R (2007). *The state of languages education in Australia: A national tragedy and an international embarrassment.* [Online]. ACSSO Australian Council of State School Organisations or the Australian Parents Council – Opinion Paper. Viewed 28 August 2009, http://www.languageseducation.com/clyneetal070328.pdf.

Díaz, A R (2013). *Developing Critical Languaculture Pedagogies in Higher Education: Theory and Practice,* Multilingual Matters, Bristol.

Ducasse, A M (2010). *Interaction in Paired Oral Proficiency Assessment in Spanish: Rater and candidate input into evidence based scale development and construct definition,* Peter Lang, Frankfurt am Main; New York.

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

Dunne, K S & Pavlyshyn, M (2012). Swings and roundabouts: Changes in language offerings at Australian universities 2005-2011. In Hajek, J, Nettelbeck, C & Woods, A (eds.), The Next Step: Introducing the Languages and Cultures Network for Australian Universities: Selected Proceedings of the Inaugural LCNAU Colloquium, 2012 Melbourne, September 26-28, 2011, pp. 9-19.

Egron-Polak, E & Hudson, R (eds.) (2010). *Internationalization of Higher Education: Global Trends, Regional Perspectives – IAU 3rd Global Survey Report,* International Association of Universities, Paris.

Fenton-Smith, B & Walkinshaw, I (2014). Research in the School of Languages and Linguistics at Griffith University. *Language Teaching,* vol. 47, pp. 404-408.

Fulcher, G (2003). *Testing Second Language Speaking,* Longman, London; New York.

Harris, A J (2013). Identifying students requiring English language support: What role can a PELA play? *Journal of Academic Language and Learning,* vol. 7, pp. A62-A78.

Heyworth, F (2013). Applications of quality management in language education. *Language Teaching,* vol. 46, pp. 281-315.

LanQua (2010). *A Quality Toolkit for Languages - Frame of Reference for Quality in Languages in Higher Education*. University of Southampton, Highfield, Southampton.

Levy, M & Steel, C (2012). The Brisbane Universities Language Alliance (BULA): A collaborative framework for university languages provision in South-East Queensland. The Next Step: Introducing the Languages and Cultures Network for Australian Universities, 2012. Office for Teaching and Learning, DEEWR, pp. 107-120.

McNamara, T (2014). 30 years on – evolution or revolution? *Language Assessment Quarterly,* vol. 11, pp. 226-232.

Michael, R (2011). Using Consensus Moderation to Ensure Appropriate Academic Standards when Marking: Self Access Standardisation Kit. *3rd Promoting Good Practice in Assessment Symposium.* Griffith University. Viewed 20 January 2013, http://app.griffith.edu.au/assessment-matters/pdfs/AcademicStandards-Marking03.pdf.

Nettelbeck, C, Byron, J, Clyne, M, Hajek, J, Lo Bianco, J & McLaren, A (2007). *Beginners' LOTE (Languages Other than English) in Australian Universities: An Audit Survey and Analysis*, Australian Academy of the Humanities, Canberra.

Normand-Marconnet, N & Lo Bianco, J (2013). Importing Language Assessment? The reception of the Common European Framework of Reference in Australia. *ECLL 2013, The European Conference on Language Learning.* Brighton England, 18-21 July 2013. Online presentation. Viewed 21 November 2013 https://www.youtube.com/watch?v=-RJfOdLnLE0.

Rust, C (2007). Towards a scholarship of assessment. *Assessment & Evaluation in Higher Education,* vol. 32, pp. 229-237.

Sadler, D R (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education,* vol. 34, pp. 807-826.

Sadler, D R (2013). Assuring academic achievement standards: from moderation to calibration. *Assessment in Education: Principles, Policy & Practice,* vol. 20, pp. 5-19.

Smith, C (2008). Building effectiveness in teaching through targeted evaluation and response: connecting evaluation to teaching improvement in higher education. *Assessment & Evaluation in Higher Education,* vol. 33, pp. 517-533.

Steed, W & Delicado Cantero, M (2014). First things first: exploring Spanish students' attitudes towards learning pronunciation in Australia. *Language Learning Journal,* pp. 1-13.

TEQSA (2011). Developing a framework for teaching and learning standards in Australian higher education and the role of TEQSA (Discussion Paper). Viewed 15 January 2013, http://www.deewr.gov.au/HigherEducation/Policy/teqsa/Documents/Teaching_Learning_Discussion_Paper.pdf.

Terrell, T D, Andrade, M, Egasse, J & Muñoz, E M (2010). *Dos Mundos*, McGraw-Hill Education, Boston.

Tripp, D (2003). Action inquiry. *Action Research e-Reports*, 017. Viewed 12 January 2013, www.fhs.usyd.edu.au/arow/arer/017.htm.

Wilson, K & Lizzio, A (2011). *Facilitating Commencing Students' Success with Early Assessment*, Australian Learning and Teaching Council, Sydney.