Evaluating the achievements and challenges in reforming a national language exam: The reform team's perspective

Carol Spöttl , Benjamin Kremmel, Franz Holzknecht
University of Innsbruck
J. Charles Alderson
Lancaster University

This paper outlines the reform of the national school-leaving exam in Austria from a teacher-designed exam to a professionally developed and standardized exam for the foreign languages English, French, Italian and Spanish, evaluating the unexpected challenges met along the way from the project team's perspective. It describes the assessment context prior to the reform to illustrate the perceived need for change and outlines the steps taken to address this need. The paper explains how key features of the exam reform project were implemented step-by-step to raise awareness with stakeholders and convince authorities to support and adopt the new approach. Reporting on the various stages of the project, it evaluates its success in introducing one standardized CEFR-based test for all students nationwide. The paper in particular highlights the unexpected political, technical and practical challenges faced, and how these were addressed, overcome or endured and with what consequences. The paper concludes with reflections and recommendations on how comparable test development projects may be approached.

Key words: examination reform, high stakes testing, challenges of innovation, politics of testing, traditional vs communicative tests

Introduction

"Changes will always be contested, championed by one set of interests over another and, ultimately, represent the triumph of particular groups and interests over others" (Buchanan & Badham, 1999, p. 171). Language examination reforms are no exception to this. East (2015) maintains that "[i]mplementing assessment reform can be challenging" (p. 101). Brindley (1998) reports that while educational authorities have introduced outcomes-based assessment in many countries, this introduction of new systems has proved problematic in some cases, mainly because of political, technical and practical reasons. Political issues, in particular, play a major role in the

Email address for correspondence: carol.spoettl@uibk.ac.at

implementation of new assessment systems (Alderson, 2009; Wall, 1996). Buchanan and Badham (1999) even claim that "the change agent who is not politically skilled will eventually fail" (p. 18). However, reform teams of language testers, despite being crucial change agents, often lack this political skill, largely because this is an underresearched area.

Literature on how change in assessment systems has been or could be managed by language testers is scarce (Alderson, 2009). Language testing is still lacking both a theory of politics in language education (Alderson, 2009) as well as a framework covering guidelines for implementing best practice models that would help reform teams manage profound and large-scale changes in assessment systems. There are some publications available that provide descriptive accounts of examination reforms (Davison, 2007; East, 2015; Mathew, 2004; Prapphal, 2008; Ramanathan, 2008). However, they often only briefly touch on the challenges faced by reform teams and rarely culminate in concrete implications from the lessons learnt or guiding lists of recommendations for future assessment reform projects.

Wall (1996) and Green and Wall (2005) are two notable exceptions to this. Green and Wall (2005) explored the political issues that influenced the work of test design teams in different military contexts. They concluded that five political issues were particularly relevant and provided valuable recommendations how to address them: (1) ownership and recognition, (2) language assessment literacy in policy-makers, (3) decision-making and information dissemination processes, (4) the role of funding bodies and (5) sustainability concerns. Wall (1996), focusing on the washback of exams on educational systems, also suggests useful guidelines. Wall hoped that her findings would provide a foundation for the construction of a framework that would help language testing teams identify potential challenges or influential factors at the beginning of their work, to enable them to predict and possibly mitigate "how they would combine to prevent the exam from having the effect that was originally hoped for" (Wall, 1996, p. 350). Twenty years later, however, this demand for a guiding framework still remains unmet. While drawing up such a comprehensive framework is beyond the remit of this paper, it does aim to add to this foundation by providing a reflexive report on the predominantly political challenges met by a reform team in the course of the complex process of setting up a standardized CEFR-based school-leaving examination for foreign languages in Austria.

The paper first briefly describes the Austrian school system, followed by a discussion of the introduction of the CEFR (Council of Europe, 2001) in the language curriculum. In Austria, as in many other European countries, the arrival of the CEFR has encouraged a move towards communicative language teaching and testing, as well as an awareness of the way standardized language exams can enhance accountability. However, although the CEFR has been central to the Austrian language curriculum in higher secondary education from 2004 onwards, the school-

leaving exam remained unchanged for the first years. The paper briefly describes the original exam, highlighting the need for change. It then outlines the exam reform process, which was initiated in 2007.

Thereafter, the paper evaluates the challenges met in the course of the reform process from the reform team's perspective. In so doing, the paper tries to raise awareness of issues that language testers are often inexperienced in dealing with. Following Brindley's (1998) and Davison's (2007) tripartite taxonomy, these will be grouped as (1) political/sociocultural, (2) technical, and (3) practical issues. The paper concludes with recommendations on how the problems encountered in these three areas could be addressed or avoided.

Local context

The Austrian school system and the introduction of the CEFR in the Austrian language curriculum

Austria has a very diverse system of secondary schooling, as shown in Table 1. Starting at age 10 (on average), children can attend four years either in a junior high school (Hauptschule), a 'new middle school' (Neue Mittelschule) or the lower grades of a higher general secondary school (Gymnasium). Schooling is compulsory until year nine (age 14), after which children can continue school in higher general secondary schools (age 15 to 18), intermediate vocational schools (age 15 to 18), or higher vocational schools (age 15 to 19), all concluding with a general school-leaving examination (Matura).

Table 1. System of secondary schooling in Austria

Age	Types of school				
10-14	Junior high school	New middle school	Higher general secondary school		
15-18/19			Higher general secondary school (15-18)	Intermediate vocational school (15-18)	Higher vocational school (15-19)

In 2004, Austria changed its national curriculum for modern foreign languages in all forms of higher secondary education. This was a direct result of the Minister of Education committing Austria to the Bologna process, which aimed at harmonizing tertiary qualifications across Europe, and to the ensuing discussion regarding the associated reforms needed in the secondary education sector. The language curriculum pre 2004 reflected long-held beliefs and traditions, both cultural and educational. As in many other European nations at the time, it was a knowledge-based curriculum, in that it outlined in detail which topics, grammatical structures and literary works were to be covered in the different grades. However, the

documentation of what topical knowledge and how much of it was deemed satisfactory lay solely in the hands of the class teachers, who were not only responsible for designing the school-leaving assessments for their own students, but also for rating them. Furthermore, the exam's focus was heavily biased towards two skills, reading and writing, with an emphasis on literary works or cultural studies topics that were memory-based.

In its place, the new curriculum laws adopted the CEFR with its focus on communicative competence and action-oriented principles. The key assessment innovation of the new curriculum was that it stipulated minimal exit level standards in terms of CEFR levels and descriptors. For the first modern foreign language (generally English), the exit level stipulated was CEFR B2, and for the second and third modern foreign languages the aim was to reach CEFR B1 by graduation, thus adhering to the plurilinguistic view of the Council of Europe. Communicative language teaching and the parity of the four skills, at least legally and in theory, had formally found their way into the Austrian foreign language classroom. For those teachers who had been practising these methods for some time, the new curriculum provided important top-down support that was long overdue, but nevertheless gratifying. The theoretical transition from a knowledge-based foreign language programme to a communicative-based one was thus initiated, as teachers were required by law to base their teaching on the principles laid out in the CEFR framework.

For the first three years, however, no thought was given to how this curriculum change might impact the school-leaving examinations. The ministry, due to a lack of assessment awareness and assessment literacy, did not anticipate the educational, political and financial ramifications of the new law. Although the new CEFR curriculum had paved the way for comparability and more transparency of both teaching and testing, many stakeholders at the time either did not comprehend or appreciate the repercussions of such a ground-breaking educational paradigm shift, or simply did not concern themselves with the consequences. As a result, the law regulating the form and procedure of the school-leaving exam remained unchanged, and teachers were still free to follow their established ways, as they themselves were still responsible for developing their own final exam. Although communicative language teaching was now anchored in the curriculum, traditional-minded teachers did not feel any pressure to change their methods of teaching or testing as long as the form of the final exam remained unchanged.

The traditional exam

The decrees regulating the school-leaving exam prior to the exam reform specified certain conditions very precisely and yet others very vaguely or not at all. Administrative issues were clearly stipulated but content issues remained inexplicit.

The dates for the exam periods were precisely stated; three different exam periods for different areas in Austria; east, west and central (presenting future standardization efforts with a challenge). The arguably excessive five-hour duration of exams was also clearly regulated (a further reform challenge for the team). In the five-hour period students were asked to write two texts, following a tradition heavily reliant on the written language skills. Little guidance was given to the class teachers (who designed the tests for their students) on topics and none on level of task difficulty. Tasks often required young candidates to solve the political problems of their day that professional politicians would have grappled with and this in a foreign language under exam conditions. Both texts had a textual prompt, which represented the reading part of the exam. However, the understanding of the textual prompt was not assessed separately, so it was possible for students to construct their written response based on their general knowledge about the topic without fully understanding the input material. Only the written production was scored (by the teacher alone), often by way of counting and penalizing grammatical errors, thus putting not the principles of communicative language learning into the foreground but rather the students' knowledge of current topics and grammatical structures. This practice therefore posed a severe threat to a meaningful interpretation of exam scores. It was also far removed from a communicative approach to language testing that would match the communicative language teaching principles as laid out in the curriculum.

The pre-reform exam was both designed and marked by the class teachers, the majority of whom had never been trained in assessment-related matters, task development or scoring practices. This jeopardized the validity, reliability and overall quality of the exam. Tasks were never piloted, seldom developed in teams or checked for their quality. In addition, over a dozen different rating scales were in use all over Austria, that is, if rating scales were used at all by teachers (Kremmel, Eberharter, Konrad, & Maurer, 2013). Some teachers used the writing rating scales of the Hungarian reform project published in the *Into Europe* series (Tankó, 2005), others used rating scales developed by commercial test publishers, while a number of provinces, schools and individuals drew up their own scales, few with professional guidance or knowledge of best practice. Training in scale use or descriptor interpretation was sporadic and variable. No standard to be measured was defined and no benchmarks or benchmarked performances were provided. The standard was simply a culturally understood and accepted notion, which differed between regions and schools.

Testing listening comprehension was only compulsory in exams for the first foreign language, and even when it was part of the test, it often followed the pattern of listening to a text and writing an essay on a topic related to the sound file. Grades were then awarded on the basis of the accuracy of the written piece, thus reducing the listening element to a mere precursor to the more important skill of writing. This

further undermined the new language curriculum's ideal of equal balance between the skills in all foreign languages.

The old exam system therefore neither met with international standards of best practice in language testing, nor did it match with the communicative principles set out in the national curriculum. The following section will outline the chronology of events of the reform process from 2007 to 2009, when the new exam was eventually legally anchored by the passing of an educational bill by the Austrian parliament in October 2009 (https://www.parlament.gv.at/PAKT/VHG/XXIV/NRSITZ/NRSITZ_00040/fname_177083.pdf).

The exam reform

The impetus for reforming the exam came from several different sectors: the teachers, the media, and the universities. There was an emerging feeling of discontent among teachers with the mismatch between the communicative curriculum and the traditional form of the school-leaving examination. This grassroots dissatisfaction was voiced at several teacher-training events across Austria. Second, fuelled by inter-provincial comparisons in the course of PISA and other standardized external test results, the debate in the media regarding the inter-provincial incomparability of school-leaving exam results gained new momentum. While public opinion had always perceived an inequality between different provinces, different schools and even between different language teachers in the same school, with the introduction of the CEFR there was now a common standard to which one could in theory compare performances. Third, university language departments, now also slowly adopting CEFR curricula, had commenced screening incoming students and were registering a considerable gap between the language level certified by the schools and the actual language level they exhibited. To some extent, the standardisation of the exam was therefore also prompted by the Bologna process (http://www.coe.int/t/dg4/highereducation/EHEA2010/BolognaPedestrians_en.asp).

The reform was initially a bottom-up approach, instigated on a very small scale. In 2005, a team from the University of Innsbruck collaborated with a group of English teachers of a local grammar school to submit a proposal to the ministry for a reformed examination entitled "A four skills Matura". Political awareness of potential legal problems arising from the mismatch between curriculum and assessment procedures was slowly emerging but general elections were imminent and the proposal found no echo and was shelved.

Crucial momentum came from another sector. Academic staff members from different institutes of the University of Innsbruck were beginning to appreciate the need for a research-based approach to a national exam reform rather than a politically imposed approach, and invited experts, national and international, to attend a round table meeting at the university. Various stakeholders in Austrian

educational politics and two expert international advisors were invited to attend. Representatives of two universities, including language acquisition and assessment experts, and the ministry's Federal Department for Innovation and Development of the Austrian Educational System (BIFIE) gathered for a discussion with headmasters and school inspectors (regional superintendents). As a result of this round table meeting, staff from the University of Innsbruck submitted a proposal to the BIFIE to establish a pilot project for a CEFR-based exam reform on a national level.

In contrast to the traditional exam developed by class teachers, the proposed exam would put the CEFR at the heart of exam development, thus aligning the new exam to the communicative language teaching principles laid out in the curriculum. The CEFR descriptors would form the basis of the test specifications for each of the skills and would also serve as benchmarks for standard setting. In addition, CEFR-based analytic rating scales with criteria from the Manual (Council of Europe, 2009) would be developed for the rating of written performances in order to break with the Austrian tradition of simply penalizing test takers for errors.

Another aim of the proposal was to develop home-grown expertise in exam development and administration based on international expertise and internationally accepted standards of best practice. For this purpose, the proposal included the involvement of an international consultant and an international trainer and a schedule for training teams of practising language teachers as item writers over the course of three years. The training aimed at providing these teachers with language testing expertise that would ensure the sustainable development of test tasks in the skills required. It further served to promote assessment literacy in Austria through the use of these item writer teachers as future pre- and in-service teacher trainers. The model also provided for the introduction of extensive trialling of test tasks.

In March 2007 the project received ministerial approval and funds were granted to the University of Innsbruck for a pilot project to develop a new CEFR-based school-leaving exam for two languages (English and French). The new exam was scheduled to be administered in 2008. The project remit stipulated that schools were not obliged to participate but were encouraged to volunteer their participation. The scope of the pilot project was restricted to the skills that, at that time, were politically low-risk but high-gain: listening and reading. Given the limited practices in place for assessing listening and reading described above, it was felt teachers would view the prospect of receiving complete test packets of these skills very favourably. Item writer training started immediately with a group of 15 handpicked general secondary school teachers from all nine Austrian provinces for English and French. The teachers were carefully chosen based on their willingness to take part in such a reform project and their roles as local disseminators, which was intended to promote positive attitudes towards the new exam among their teacher colleagues and students.

Training began in May 2007 and by November sufficient reading and listening tasks had gone through two sets of trials to allow the first standard setting session to be conducted in December of the same year. Stakeholders invited to the session included teachers, headmasters, school inspectors, teacher trainers, university staff, BIFIE staff and ministry representatives and one external international testing expert. The tasks approved in this meeting made up the test booklets for the first live administration of this new part of the exam for 56 pilot schools in five of Austria's nine provinces in May 2008. The positive response to this administration by the participating pilot schools and the stakeholders involved in standard setting led the ministry to approve further funding.

From 2009 to 2010, the project expanded and two new teams of item writers were required for task development. One team targeted the receptive skills and language in use (a test paper assessing lexico-grammatical knowledge) in two new languages, Spanish and Italian. The second team was required to develop writing tasks to pilot the inclusion of standardised tests of writing in the English and French exams. This team was tasked with the job of developing national assessment scales at two CEFR levels in two languages, English and German (Holzknecht et al., forthcoming; Konzett, 2011). The first standardised writing tasks were trialled on a small scale in May and September 2009 and in January 2010 on a larger population nationally. The scales were developed over a three-year period. By 2010, test tasks for listening, reading, language in use and writing were developed centrally by item writer teams, for four languages and two CEFR levels as outlined in Table 2 below

Table 2. Skills and CEFR levels of the new exam (6y and 4y relates to the number of years students have been learning the second foreign languages French, Italian and Spanish, which varies from school to school. English, as the first foreign language, is taught already in primary school.)

	Listening	Reading (6y)	Reading (4y)	Language in Use (6y)	Language in Use (4y)	Writing
English	B2	B2		B2		B2
French	B1	B2	B1	B2	B1	B1
Italian	B1	B2	B1	B2	B1	B1
Spanish	B1	B2	B1	B2	B1	B1

In addition to item writers, additional staff was needed to handle the logistics of test development, such as database management, organisation and administration of field trials, data entry, statistical analyses of trial data, and workshop preparation and assistance. The staff for these tasks were recruited among students at the University of Innsbruck. The language teacher training curriculum at the University of Innsbruck was unique at the time in that it was the only one across Austria which included a compulsory module on testing and assessment, established in 2004. The project leader, who was teaching this module, was able to identify potential staff, i.e. language students who showed interest in the area of assessment and who were committed to their studies. The students also had the opportunity to write their

Master thesis on an assessment-related topic, supervised by the project leader. This allowed action-based research to be conducted concurrent to the test development process. For example, one Master thesis drew up a needs analysis for Austrian English teachers in developing and assessing writing tests (Holzknecht, 2009). The needs determined by this study formed the basis of teacher training courses subsequently implemented by teacher training institutions across Austria. Conducting this kind of research also ensured that the challenge of identifying and training suitable staff was sustainably met.

In October 2009 the government finally passed a new educational bill, anchoring the exam reform legally. But parliaments do not pass laws for single subjects such as foreign languages. This fact caused a number of unanticipated changes and made the bill more far-reaching than originally envisioned. Firstly, the bill encompassed all subjects. The CEFR-based exam development had won favour with politicians, who were in favour of the idea of measuring students' achievement on a common, clearly defined standard. They deemed such a competence-based approach both politically desirable and transferable to all other subjects, core and elective alike. This has since meant that subjects like Mathematics, German as L1, and the sciences have followed suit in defining core competences and levels of achievement, similar to the CEFR. Secondly, the new bill determined that selected core subjects must be standardized, requiring all students across Austria to be given the same exam questions. It also necessitated a third major change: the previous system of three exam dates east, west and central, was deemed no longer viable (largely for financial reasons) and one national date was prescribed with all regions offering the same subject on the same day. Finally, an extremely controversial but audacious step was taken to offer one standardized exam in the foreign languages across all school types: higher general secondary and higher vocational secondary schools. As a consequence, the impact of this reform was to be felt by around 15,000 students in higher general secondary schools and 22,000 students in higher vocational secondary schools.

Although the project was funded and overseen by the BIFIE, at the beginning all stages to set up the standardized language exam were carried out by the team at the University of Innsbruck in collaboration with the international trainer and the consultant. This work included item writer training, item moderation, organization and administration of field trials, statistical analyses of field trials, and standard setting. The BIFIE only took on a supervisory role. However, the BIFIE's influence over the individual areas in the test development cycle grew in parallel to the project's size. This was also due to the fact that the university project team would be mainly responsible for the initial development phase of the exam reform, but responsibilities would transfer later on as the BIFIE would take over the routine production phase after the development project had ended. Thus, the success of the pilot project also strengthened the BIFIE as an institution. In the later stages of the project, BIFIE also took over completely the model set up for field trials and standard

setting established by the team in Innsbruck. They also improved on the models for the delivery of the live administration, which could not have been done sustainably by the Innsbruck team as it requires an institution with more appropriate legal and political standing and decision-making powers to enforce such changes within the system. Needless to say, however, the work across two organisations, particularly in phases of transition from the development phase to the production phase, brought its own challenges, mainly relating to communication and negotiating areas of responsibility.

The BIFIE's long-term goal was to fully institutionalize the new exam and move all developmental processes to the centre in Vienna. Once the project reached a certain size, team members from Innsbruck were offered BIFIE contracts, and all of the logistics were moved from Innsbruck to Vienna step by step. By 2012, the majority of the Innsbruck team migrated to Vienna to work for the BIFIE full time. Together with the project leader, the remaining members of the original team established a Language Testing Research Group at the University of Innsbruck.

Unexpected Challenges

Eckes et al. (2005) state that "[g]iven the considerable cultural diversity of European countries it should not come as a surprise to see that each country […] has faced unique problems of language teaching and assessment and has developed specific solutions to these problems" (p. 356). This section describes the unexpected problems faced in the Austrian exam reform process and how these were addressed. Although these problems were unique to the Austrian situation at the time, language testers in different national contexts who attempt similar reforms might also encounter many of them. In the following discussion, these challenges have been categorised as (1) political/sociocultural, (2) technical and (3) practical issues, according to the taxonomy provided by Brindley (1998) and Davison (2007).

1. Political/sociocultural challenges

Lack of external assessment culture and fear of change

Austria does not have a tradition of external assessment. Similar to its neighbouring countries Slovenia and Hungary, the form and role of the school-leaving examination dates back to the Habsburg monarchy (Pižorn & Nagy, 2009). The Austrian education system, including the language education system, had always been heavily knowledge-based and content-oriented. The arrival of the CEFR, with its focus on communicative competence rather than on topical and grammatical knowledge, thus implied a radical paradigm change. In addition, there was no culture of accountability in education. Teachers, despite having received hardly any pre- or in-service training in assessment, were considered to know best by default.

Local school inspectorates generally checked the content of Matura exam tasks in advance, but since trained testers were not involved, these checks added little to the quality of the exam. There was no systematic approach or structure to hold teachers accountable for their work. Therefore, paradoxically, there was simultaneously both a strong feeling of dissatisfaction with the present assessment system as well as a fear of the unknown that any exam reform would bring.

With the advent of a new assessment system for this most high-stakes of national exams came numerous uncertainties that translated into a fear of change at several levels in the system. Despite a general national mentality of reluctance towards reforms, the number of worries of different stakeholder groups that surfaced in the early stages of the project was unexpected. Teachers publically voiced fears that educational standards were dropping. The new transparency, although often listed as an argument for setting up such new exams, was thus at the same time one of the biggest dangers for the success of the new exam. The pending, and inevitable, comparisons within schools and provinces were looming large in stakeholders' minds. Behind this were concerns about likely weaknesses in teachers' language competences and teaching practices being exposed through the exam results, which could eventually lead to potential negative repercussions for individuals and schools. The new exam was also perceived as additional work and as showing a lack of trust in teachers' competences as well as a threat to their status and powers as teachers, testers and gatekeepers. This "painful unclarity" (Goodlad, Klein, & associates, 1970) of stakeholders being called upon to implement changes that they do not fully understand inevitably creates confusion, frustration, and anxiety.

Hidden agendas

Pižorn and Nagy (2009) are convinced that "the most important element of any reform project are the individuals and their ambitions, personal agendas, openness to change and attitudes to professionalism, in short, micropolitics" (p. 185). Evidence of this was also found in Austria. While some of these agendas were relatively obvious and predictable, some were more opaque, especially for the reform team who were newcomers to the political stage. Openness to or, as discussed above, fear of change also played a crucial role in the agendas of the different stakeholders. Hidden agendas presented another unexpected challenge.

The agendas of political factions are generally predictable. These are usually strongly driven by their ideology and political competition, resulting mostly in debates about the costs, organization and administration of the exam rather than any underlying testing principle. For example, opposition parties seized on any slip-up to attack the education minister in charge, aiming to weaken the minister and hoping for headline coverage.

The strong opposition from teacher unions was less expected. Teacher unions are traditionally very strong in Austria and closely linked to the big political parties, particularly the Conservative party. From the reform team's perspective, they offered strong opposition to the exam reform for ideological and party political reasons on issues regarding loss of status and power. Teachers and their unions strongly opposed the new exam due to an anticipated loss of income. Even though teachers welcomed the reduced workload implied by a standardized exam, they did not want to lose the additional money they traditionally received for administering and marking the Matura and therefore many opposed a completely externally designed, administered and marked standardized exam. This was one of many reasons for the fact that marking, absurdly, remained with the class teachers (see *Quality control – reliability issues* below for more detail).

Universities and their representatives had a fairly obvious agenda, i.e. their budget. They anticipated a cost reduction in the language faculties through streamlining the admission procedure to their language courses, and excluding students below the required entry level which the new exam would facilitate. However, the teacher training colleges' interest in the reform was that they would be delivering the training courses needed and called for by the (mostly in-service) teachers. They welcomed the increased funding the government began to allocate as of 2009.

School inspectorates and their legal departments were mostly concerned about appeals and any legal aspects that would threaten the smooth administration of the exam. Discussions about cut scores, for example, were particularly sensitive.

Student representatives at individual schools are usually elected for a one-year period. Student representative committees were therefore interested in short-term goals. Their role and agenda was unexpected as they were mostly campaigning for a postponement of the exam to buy students more time for exam preparation by arguing that there were still too many uncertainties to be resolved. Eventually, these representatives, with the support of the parents' associations, managed to push back the first compulsory live administration date of the Matura by one year. In addition, it was often observed just how easy it was to motivate this stakeholder group to oppose the new exam.

Such political agendas frequently frustrate test developers as they often overrule empirically-grounded proposals, and might therefore impact the quality of the exam, as was the case in Austria. Directors of the BIFIE, for example, were frequently publically criticized, dismissed and replaced. Frequent rotation of personnel was common at several institutions, including parent teacher associations who were only interested in the exam year of their own child. This discontinuity made work difficult for the research team, because new players seldom brought much assessment literacy to the table, but often new personal agendas.

Dependence on other subjects in political decisions

Another political challenge was the dependence on other subjects in terms of political decisions. The authorities, with their lack of understanding of language testing, lumped all subjects together when it came to decisions about, for instance, test administration. The lack of a standardized speaking part in the exam is one example for this, which will be illustrated in the following.

The original aim of the project was a four-skills exam, testing reading, listening, writing and speaking. The project leader had paved the way for this in numerous teacher training sessions across Austria, the majority on assessing speaking, well before the project had received government funding, and the project had, in fact, developed a degree of assessment literacy among teachers in terms of speaking assessment, including a more accurate understanding of the construct of speaking, rating procedures, and interlocutor roles. Inroads had also been made into designing speaking tasks with an awareness and acceptance of specific, more authentic task types than those traditionally used.

However, the overriding political goal was to establish a competence-based exam that would span all subjects, particularly the major subjects German, Mathematics and the foreign languages. Negotiations at the Ministry level therefore necessitated compromise. Political decisions meant that a one-fit model was required for all subjects with no exceptions, however sensible or obvious the need for an exception was. The model the Ministry was prepared to support was one written exam for all subjects that was standardized. Although this allowed for the opportunity to have an oral exam for all subjects, such oral exams had to remain the responsibility of the class teacher. At the same time it was necessary to ensure that the same parameters applied to all oral exams. This translated into such odd regulations that an oral exam in Maths had to be the same length as an oral exam in a foreign language, even for second foreign languages where the target proficiency level meant it was hardly possible for candidates to fill the speaking time allocated. However, any questioning of this political decision could have jeopardised the entire reform project and risked any reform progress at all.

In the interim period, the reform team managed to soften the effect of this. Although there are still no standardized speaking tasks across Austria for the oral part of the school-leaving exam, the project was successful in providing teachers with sample tasks, analytic and holistic rating scales, rating forms and some benchmarked performances.

2. Technical challenges

Quality control – external validation

External evaluation or validation is an essential phase of any test development process (Downing & Haladyna, 2006). Accordingly, the original model and budget for the Austrian exam reform project had included this phase and had already identified and approached external consultants to carry out such a validation study. Parallel to this, however, the authorities hoped that a professional non-commercial testing body or organization would offer to provide a service to ministries and institutions of approving CEFR-linked exams or certifying exams claiming CEFR-linkage. While discussions of this idea had begun at various testing conferences and even resulted in a task force to debate this function (e.g. http://www.ealta.eu.org/documents/archive/agm2013.pdf), no organization has yet decided to provide this service.

Since the BIFIE had in the meantime also created an internal department for evaluation, the urgency of an external evaluation disappeared as authorities considered the need for evaluation and quality control covered. The need for external evaluation was believed by BIFIE to be mitigated by applying for and being awarded EALTA institutional membership, which supposedly certifies adherence to best practice in test development. An external evaluation of the exam is therefore still pending, but although it is not out of the question, it does not seem a priority for administrators at this moment. However, a change of key players at the ministry level has meant that this is under review at the time of writing.

Quality control – reliability issues

Although the project was successful in establishing central marking by item writers for all open-ended listening, reading and language in use items after piloting, it has not yet succeeded in setting up a national model for central marking for the live exam. The marking of all sections of the live tests is still the responsibility of individual class teachers. There are various reasons for this, some stemming from pressure from sectors of the educational community and others from various specific legal restrictions.

The group closest to exam marking are teachers. In the initial stages of the project, there was clear fear on the teachers' part that they would lose control over the final grades of their students. This fear was heightened by the prospect of transparency of scores, feeding the concerns mentioned above that low grades might unfairly put teachers in a bad light on a national scale by disregarding the fact that quality of teaching is only one of the factors that contributes to a student's performance. Opposition was fuelled by a belief that teachers themselves would be assessed by the new reforms. In addition, the great majority of teachers did not see the need for

central marking, as they felt that they were trained professionals who could and should do the job. Teacher unions and school inspectors shared this view. However, as testing had not generally been part of the teacher training curriculum, and seldom involved rater training, the great majority of teachers were not aware of ethical issues in marking, such as inter- and intra-rater reliability, and impartiality to test takers.

Another reason for failing to establish central marking had to do with the associated costs for Austria, and above all, a lack of political will. Although calculations showed that in the long run it may not have been more costly to set up central marking than the current system of paying each individual teacher, training a sufficient number of professional raters to mark test takers' exams would have overstretched the resources available. Moreover, central marking would have required another change of the law, as according to the curriculum then in force, students should receive their marks two weeks after the final exam, a timeframe which was not realistic in a central marking scenario. However, the Ministry felt that any additional legal changes might have unduly threatened the project as a whole. This fear, together with the strong opposition from teachers and unions, prevented the ministry from pursuing the idea further.

3. Practical challenges

Media

The break with the previous assessment tradition, coupled with a general uncertainty and fear of change as outlined above, was picked up quickly by the national media. A lack of language assessment literacy and more importantly the frequent unwillingness or inability to understand the procedures, motivations and benefits behind the new exam on the part of reporters and journalists were completely underestimated by the project team.

The project leader had to give a number of press interviews that required her to explain the issues and complexities of standardized testing in a very limited amount of time (a time span that was dictated by the journalist, dependent on when she had to go to press). This was problematic enough were it not for the fact of the unexpected challenge of the slant of each article. Journalists tended to select their own angle from the data gathered in the interview and often it was too late for the reform team to realize that this was based on factors outside the interview topics, an incomplete understanding of the issues at stake, or the fact that journalists reported only on those issues they (thought they) understood.

Although all parliamentary parties generally supported the idea of a standardized school-leaving exam, it nevertheless was constantly used as a political football, with all setbacks and lapses publicly exploited by competing parties in the media. Cut scores present a pertinent example here. For commercial exam bodies, setting the cut

score is an internal decision. In a national exam where teachers correct the exam papers, setting the cut score has to be publically explained. When, for example, the cut score was adjusted from one year to the next based on results from piloting and standard setting, thus ensuring exam equivalence and fairness, both national and local newspapers were extremely keen to portray this as a flaw in the exam, as chaotic procedures and as evidence that the exam designers did not know what they were doing, when, in reality, the standard procedures of international best practice were being followed. Moreover, although different teams were responsible for different exam subjects (mathematics, German, the sciences, and the modern foreign languages), failures of one team were swiftly generalized and dismissed as failures of the entire exam development project. The exam reform team realised too late that such a major exam reform at a high-stakes and national level required not only research-based professional development but also professional "marketing" to proactively involve the press.

The cost of extensive monitoring research

Ideally, newly developed high stakes testing instruments such as the standardized Austrian exam would conduct extensive monitoring research before they are fully implemented. In many scenarios, however, monitoring is not introduced until after the exam is put into operation (Alderson, Clapham, & Wall, 1995), or in some instances is not carried out at all. This is also the case of the Austrian Matura, which has not been extensively monitored to date. Although international experts have been consulted throughout the development process and have overseen the development and setting up of the exam by implementing EALTA's guidelines of best practice, full-scale monitoring studies are still lacking. The BIFIE have yet to address this issue. However, as Fullan (1991) points out, it may take a long time before any such innovation really "takes hold" (p. 351). Thus, the very fact that such an institute has been installed and the authority has invested in national structures as a basis for further improvements is in itself an achievement in managing change, even if the capacity of such institutions in the short term is not inspiring. Alternatively, it could be argued that this is the successful start of an ongoing process of change in the right direction and that, despite short-term frustrations, the conditions for further improvements are set up.

Recommendations for exam reform

The reform of a national high-stakes exam is clearly a complex endeavour. The technical aspects of test development that language testers are all too familiar with are only one source of complexity. This article has tried to raise awareness of the numerous challenges on the micro and macro political level that accompany such a test reform project and that many language testers are less experienced in dealing

with. The lessons learned in dealing with the challenges and shortcomings outlined above lead us to propose the following nine recommendations for testers who are about to take on a project of this dimension. These are again clustered according to the taxonomy introduced earlier:

Political dimension

1. **Ensure sufficient and sustainable resources for training to increase (and maintain) language assessment literacy.** Addressing assessment literacy at a national level requires training and educating in assessment-related matters, but it needs to be understood that this is a never-ending process. It requires appropriate anchoring in pre- and in-service programmes. "Teacher in-service training is key in making changes in examination systems" (Ramanathan, 2008, p. 124). Training programmes should accompany the exam reform throughout and cannot be done in one-off events. It takes considerable time to reach most teachers, by which point many participants of the first training courses will be in need of a refresher course or will require updated information as the exam reform evolves. Moreover, given that staff turnover is inevitable, personnel in all institutions involved also require constant training to ensure both continuity and sustainability. Enhancing the language assessment literacy of all parties involved may be the only way to address or prevent shortcomings such as the ones described above.
2. **Consider involving stakeholders in roll-out events.** Preparing information events jointly on new developments that affect teaching and classroom assessment is one way of assuring a deeper understanding of the issues involved. Test developers need to be aware that changing the technical features of an exam is only a small part of exam reform. Changes in the educational system need to go hand in hand with the technical aspects of test development. This means that exam reform cannot be introduced without involving, and educating, all stakeholders.
3. **Keep your finger on the political pulse and keep in touch with the key players in the educational system.** Test development and test research can never be detached from politics and financial resources. Governments come and go as do ministers, inspectors and reporters. Keeping abreast of these developments and stakeholder representatives with their overt and hidden agendas is a key to successful project management.
4. **Be aware of hidden agendas.** Individuals may have very different motives for their involvement in the project. These motives may be laudable and altruistic with individuals supporting or opposing the exam reform for very different and highly individual reasons. Although this micro-political dimension of assessment reform is inevitable, an awareness of it can avoid potential blows to the motivation of the reform team. Test developers need to be aware that politics (and

individual political players) can influence their work and stakeholders might hijack a project for their own causes.

Technical dimension

5. **Document the processes, issues, challenges and threats carefully throughout the exam reform project.** Although the documentation of the technical aspects of assessment reform is generally acknowledged as best practice, the documentation of the social and political dimensions of such projects is often a neglected area. Frequently this is simply because test development teams are fully engaged with the development, trialling and administration of the exam. Yet, having a designated documentation officer for the test development project would enable a wider circle of people to reconstruct the genesis of policies, decisions and the reasoning behind them. This offers various additional benefits for new staff joining the project and safeguarding against them reinventing the wheel, or even turning the wheel backwards. For the original team it offers a resource for reflection and evaluation after the end of the project.
6. **Be prepared for compromise.** Initial plans of the ideal exam reform are unlikely to be fully implemented for financial or political reasons or both. Be prepared to negotiate and accommodate, as it may be better to make short-term compromises on some elements than to have no exam reform at all. Compromise may even require relaxation of key test requirements like rater-reliability. In the project described here, an unwillingness to relinquish teacher control was one considerable obstacle to central correction that led to standardised tasks being graded by the classroom teacher. These same teachers are now calling for a system-central correction. Make sure to document the gap between goals and subsequent compromises so that the long-term aims of the project are at least potentially kept in view.

Practical dimension

7. **Engage professional help in handling the media.** Reforming public exams impacts on a wide range of stakeholders from pupils and their families to teachers and their unions, from school heads and their inspectors to ministers and the press. Issues of wide public interest will cause interest and controversy and, in education, everyone is inclined to regard themselves as an expert. Additionally, the media has a tendency to try to find assessment scandals even where there are none. Effective dissemination of information is key to success in addressing the fear of change. Educating and answering journalists needs to be recognised as a crucial and an on-going process. Given that this may not be familiar terrain for testers and is seldom, if ever, addressed in academic conferences, communication with the press needs to be managed, proactively, by professionals.
8. **Establish a chain of command for information release.** Information is power and untimely release of information can adversely affect public perceptions of an

exam. Even though people involved in the test development might be happy to share what kind of test formats may be used with an interested journalist, or item writer teachers may want to discuss with their peers what kind of assessment criteria might be incorporated into a rating scale, making such information public too soon can cause more confusion and insecurity than is necessary or desirable, adding fuel to the fire for opponents of the reform. Developing an awareness among those involved in the test development process of potential negative repercussions regarding new developments is beneficial, and determining a hierarchy of information release of who is allowed to say what and when about an exam, could well prevent any latent backlash.

9. **Establish a clear division in areas of competence when working across institutions and anchor this contractually for the duration of the project.** Although working with different institutions can be beneficial in terms of developing synergies, responsibilities need to be clearly laid out, if possible, from the start. Working across different institutions can have numerous advantages and drawbacks, but it makes for a stronger and more productive partnership if the institution in charge of test development has the power to decide on testing matters and if those partners with the political remit tackle the politics required to deliver these decisions.

Conclusion

This paper outlined the reform of the national school-leaving exam for foreign languages in Austria. It described the transition process from a classroom teacher-designed exam to a professionally developed and standardized exam, evaluating the unexpected challenges met along the way. The challenges and recommendations on how these could be addressed or avoided were presented from the project team's perspective. Following Brindley's (1998) and Davison's (2007) tripartite taxonomy, they were grouped as (1) political/sociocultural, (2) technical, and (3) practical issues.

Although the Brindley/Davison framework has proven useful initially to systematically evaluate and present the challenges, it appears important to point out a major limitation of this framework at this point. In the innovation process of such a reform project, the theory and technicality of testing meets the real world, the politics and practicalities of which permeate all aspects of the process. As a result, several factors are at play in any given challenge faced, and a categorization into either political or technical or practical may be impossible or, at the very least, fall short of fully accounting for the issue. While we have decided to categorise our experiences in the way they are presented here, this does not mean that they could not also fall under either of the other two categories. However, a discussion or adaptation of the Brindley/Davison taxonomy was beyond the scope of this paper.

Despite the shortcoming of the taxonomy in capturing the interplay between different factors, this paper has tried to raise awareness of the numerous challenges on the micro and macro political level that accompany a test reform project and that many language testers are less experienced in dealing with. It further hopes to contribute to an increased understanding of the kind of tension language testers often face between the necessity to enter and engage in the political fray while at the same time wanting or needing to remain distant from or untinged by it. However, managing change has a political dimension almost by definition and language testers should be aware that changing an assessment programme is complex, but is a comparatively minor challenge compared with the necessary changing of mind-sets. Developing assessment literacy is a long process that takes passion, persistence, patience and political skill.

References

Alderson, J. C. (2009). *The politics of language education: Individuals and institutions*. Bristol: Multilingual Matters.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing*, *15*, 45–85.

Buchanan, D., & Badham, R. (1999). Power, politics and organisational change: Winning the turf game. London: Sage Publications.

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

Council of Europe. (2009). Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR). Strasbourg: Language Policy Division.

Davison, C. (2007). Views from the chalkface: English language school-based assessment in Hong Kong. *Language Assessment Quarterly*, *4*(1), 37–68. http://doi.org/10.1080/15434300701348359

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. New Jersey: Lawrence Erlbaum Associates.

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing*, *32*(1), 101–120. http://doi.org/10.1177/0265532214544393

Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & Tsagari, C. (2005). Progress and problems in reforming public language examinations in

Europe: cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. *Language Testing*, *22*(3), 355–377.

Fullan, M. (1991). *The new meaning of educational change*. New York: Teachers' College Press.

Goodlad, J. I., Klein, M. F., & associates. (1970). *Behind the classroom door*. Columbus, Ohio: Charles Jones.

Green, R., & Wall, D. (2005). Language testing in the military: Problems, politics and progress. *Language Testing*, *22*(3), 379–398. http://doi.org/10.1191/0265532205lt314oa

Holzknecht, F. (2009). A needs-analysis for Austrian English teachers in developing and assessing writing tests for the Matura. Unpublished MA dissertation, University of Innsbruck.

Holzknecht, F., Kremmel, B., Konzett, C., Eberharter, K., Konrad, E., & Spöttl, C. (forthcoming). Potentials and challenges of teacher involvement in rating scale design for high-stakes exams. In D. Xerri & P. Vella Briffa (Eds.), *Teacher involvement in high stakes language testing*. Springer.

Konzett, C. (2011). *Every word counts. Fine-tuning the language of assessment scales: A field report*. Paper presented at the IATEFL TEASIG conference: "Standards and standardizing in high and low stakes exams. Assessment from classroom to Matura.", Innsbruck.

Kremmel, B., Eberharter, K., Konrad, E., & Maurer, M. (2013). *Righting writing practices: The impact of exam reform*. Poster presented at the 10th annual EALTA conference, Istanbul.

Mathew, R. (2004). Stakeholder involvement in language assessment: Does it improve ethicality? *Language Assessment Quarterly*, *1*(2-3), 123–135. http://doi.org/10.1080/15434303.2004.9671780

Pižorn, K., & Nagy, E. (2009). The politics of examination reform in central Europe. In J. C. Alderson (Ed.), *The politics of language education: Individuals and institutions* (pp. 185–202). Bristol: Multilingual Matters.

Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing*, *25*(1), 127–143. http://doi.org/10.1177/0265532207083748

Ramanathan, H. (2008). Testing of English in India: A developing concept. *Language Testing*, *25*(1), 111–126. http://doi.org/10.1177/0265532207083747

Tankó, G. (2005). *Into Europe - the writing handbook*. Teleki Lazlo Foundation and The British Council Hungary.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing*, *13*(3), 334–354.