International language proficiency standards in the local context: Interpreting the CEFR in standard setting for exam reform in Luxembourg

Abstract

In the field of second and foreign language learning, the Common European Framework of Reference for languages (CEFR; Council of Europe, 2001) is widely-used for setting language proficiency standards within European, and increasingly global, contexts (Figueras & Noijons, 2009; Martyniuk, 2011). Few studies, however, have investigated the ways in which systemic, macro-level factors within national educational contexts may influence standard setting practices using the CEFR. In this paper, we explore this issue through an analysis of recorded discussions within standard setting sessions for the *Épreuve Commune for English*, a national English language examination in Luxembourg. The data reveals four key sources of influence on standard setting decision-making: Luxembourg's unique language ecology, streamed schooling, the national curriculum, and an ongoing exam reform project. Through this analysis, we argue that Luxembourg functions as a critical case illustrating the tension between international standards of language proficiency and local realities.

Background

In many educational systems around the world, qualifications frameworks have been adopted to provide comparability of qualifications across awarding bodies within and between countries, and greater transparency for stakeholders (e.g., employers, educators) in interpreting the competencies of a qualification holder (Johnson & Wolf, 2009). Qualifications frameworks are also argued to be tools for educational policy evaluations and

reforms (Young & Allais, 2013), given that they are often implemented in a top-down manner.

While in principle this should enable transparency from the individual to the global level, research and experiences from several countries have shown that the implementation and application of frameworks within and across contexts is not without its challenges and does not automatically lead to greater comparative understanding of competencies (see e.g. the chapters in Young & Allais, 2013; Allais, 2014; or the 2009 Special Issue 'The challenge of cross-border qualifications recognition' of the journal *Assessment in Education: Principles, Policy & Practice*).

Frameworks in language education: The CEFR

In the field of language testing and assessment a key framework is the Common European Framework of Reference for Languages (CEFR), published by the Council of Europe (2001).[1] Among other aims, the CEFR was developed to enable 'a common basis' for, and transparency in interpretation of, language curricula, teaching materials, evaluations, qualifications, etc. across contexts (e.g., educational, employment, geographical contexts, etc.) (Council of Europe, 2001, p.1). It promotes a communicative approach to language learning, with emphasis on the 'knowledge and skills [language learners] have to develop so as to be able to act effectively' (Council of Europe, 2001, p.1). The Framework provides sets of illustrative descriptors, also called 'common reference levels' (Council of Europe, 2001, p.22), of what a language learner can do at each of six levels of language proficiency. From lowest to highest, these are labelled: level A1, A2, B1, B2, C1, and C2. An example of a descriptor for listening ability at the A2 level – the level under focus in the current study – is

[1] Note that in 2018 a Companion Volume was published by the Council of Europe, filling gaps in the original scales, adding new scales/descriptors for e.g. mediation, phonological control, sign languages, young learners, and removing reference to the 'native speaker' as a reference point.

that a language learner who has reached this level of listening proficiency '[c]an understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated' (Council of Europe, 2001, p. 66).

The CEFR has been widely adopted in language learning, teaching, and assessment across Europe, and, over the years, also in wider international contexts (e.g., see chapters in Figueras & Noijons, 2009; Martyniuk, 2011). For example, language courses and textbooks are labelled as targeted at a particular CEFR level, and specific CEFR levels are set as language proficiency requirements for entry into higher education, for visa and migration purposes, and for professional certification. With reference to assessment, large numbers of commercial and non-profit organisations, as well as educational institutions, now describe their language tests and report test results in terms of CEFR levels (e.g., the Cambridge English Qualifications, TOEFL iBT, IELTS, or school-leaving language exams such as in Austria or Slovenia). In order for such links with the CEFR to be deemed valid, it is necessary to first conduct a formal process of alignment.

One widely-used approach to formal alignment is through standard setting. As Tannenbaum and Wylie (2008, p. 2) explain, '[s]tandard setting is a general label for a number of approaches commonly used to identify test scores that support decisions about test takers' […] level of knowledge, skill, proficiency, mastery, or readiness.' It involves a group of people with relevant expertise who make judgments, in a principled manner, on the extent to which a test item or performance reflects the characteristics of a certain level descriptor. This information then feeds into decisions on cut scores, or, in the context of the CEFR, on links with specific CEFR-levels (see e.g. Cizek and Bunch, 2007).

The practice of aligning language assessments to the CEFR through standard setting procedures is now commonplace, both among commercial testing providers (e.g., see

Brunfaut & Harding, 2014; Lim, Geranpayeh, Khalifa & Buckendahl, 2013; Papageorgiou, 2009; Tannenbaum & Wylie, 2008) and for national examinations (e.g., Spöttl, Kremmel, Holzknecht & Alderson, 2016). Such practices are routinely guided by the *Manual* (Council of Europe, 2009) – a text produced specifically to provide practitioners with methods and techniques for carrying out standard setting within their own contexts. The *Manual* contains advice on how to describe information concerning the exam, the phases that need to be completed as part of the linking process (familiarisation, specification, standardisation training, standard setting, and validation), and how to report supporting evidence on the linking endeavour. Numerous published CEFR-alignment studies have followed the specific standard setting methods outlined in the *Manual* (e.g., Brunfaut & Harding, 2014; Papageorgiou, 2009; Tannenbaum & Wylie, 2008).

Even though the CEFR has become a globally-influential framework, language assessment scholars have critiqued the document on the grounds that it often functions as a top-down set of standards imposed by policy-makers, and that it lacks a strong theoretical basis in second language acquisition (for discussion of this debate see Deygers, Zeidler, Vilcu & Carlsen, 2018). Questions have also been raised concerning the interpretation of the CEFR levels across contexts (e.g., Holzknecht, Huhta, & Lamprianou, 2018; Lim, Geranpayeh, Khalifa, & Buckendahl, 2012), a critique which has direct implications for standard setting. The CEFR is, to some extent, intended as a 'context-free' framework. Its claim to work across different languages and regions is premised on its malleability. As Milanovich and Weir (2010) have explained, 'the CEFR itself is deliberately underspecified and incomplete … It is precisely this feature which makes it an appropriate tool for comparison of practices across many different contexts in Europe and beyond … [Users must] adapt it flexibly to suit local purposes' (p. x).

There are few studies, however, which have empirically investigated the ways in which such local adaptation is achieved in practice, or what the systemic or macro-level factors are within specific contexts which may influence the way in which standards are set using the CEFR framework. This issue is important as the 'messiness' of standard setting may often be masked by the technical reports produced, while the validity of standard setting itself requires transparency. Understanding such systemic issues will contribute to a more robust theory of standard setting, helping to make clear the implications of situations where international standards collide with contextual realities.

This article seeks to address this issue, and will, in practice, explore it within the context of Luxembourg, by analysing the macro-level, systemic influences in a standard setting session for a national English language examination: the *Épreuve Commune for English*.

Luxembourg and the Épreuve Commune

The Grand Duchy of Luxembourg is a small country (2,586 sq kilometres; 590,000 inhabitants – 1 January 2017; Visitluxembourg.com, n.d.) in Western Europe. Its national language is Luxembourgish, while its administrative and judiciary languages additionally include French and German. Furthermore, having become a 'country of immigration' (Luxembourg.lu, n.d.), with 46% of the population being non-Luxembourg nationals (e.g. 16% Portuguese), several more languages are used within Luxembourg society.

The country's plurilingual policy is reflected in its state education system, in which children start their pre-school education through the medium of Luxembourgish, and then switch to German as the medium of instruction during primary and lower-secondary school. French is taught as a subject at these levels, and used as the medium of instruction for mathematics teaching at lower-secondary school. English is also introduced as a mandatory

subject from the second year of secondary school. Finally, for those learners in the so-called 'classical' stream (oriented towards higher education), the medium of instruction changes to French in upper-secondary school (Men.lu, n.d). This means that 85% of 15 year olds are taught through a medium that is different from their home language (e.g., 45% Luxembourgish speakers, 22% Portuguese speakers) (European Commission/EACEA/Eurydice, 2017). On the basis of this policy, compared with other countries in Europe Luxembourg has among the largest numbers of different languages taught in general education, as well as one of the highest rates of classroom hours allocated to language teaching.

With respect to English language teaching, until 2007, so-called traditional grammar- and literature-oriented approaches prevailed in Luxembourg schools. In fact, Geyer (2009, p.1) states that English was taught 'as a "truly" foreign language', despite being 'widely spoken' within Luxembourg (Luxembourg.lu, n.d.). From 2007 onwards, however, driven by major educational reforms, the approach to language teaching and learning began to shift to a competence-based, communicative, CEFR-aligned approach (MEN, 2010, 2011a, 2011b). The curriculum revisions were gradually phased-in, starting with the lower years of secondary education (MEN, n.d.). Then, in 2011, in response to calls from teachers who were concerned about an increasing gap between how English was taught (the new competence-based approach) versus how learners were assessed (still a more traditional approach), a project was introduced to implement the competence-based approach in the assessment of English at the secondary school level (see Brunfaut & Harding, 2018 for more information). The project was supported by the Ministry of Education and involved a group of English teachers from Luxembourg (the Test Design and Evaluation [TDE] team) working with external language testing consultants from Lancaster University. The TDE team, together with the consultants, decided as a first step to develop a standardised English test for lower-

secondary school. A national test already existed for French and German, called the *Épreuve Commune,* which aimed to evaluate learners' achievement of the curriculum targets at the end of the second or third year of secondary school. No such test existed for English, however, and thus the team (in agreement with the Ministry) started to work on an *Épreuve Commune for English*.[2] A key aim of the new test was to bridge the gap between the curriculum and assessment. Since the basic target for English at lower-secondary school was defined by the curriculum as CEFR level A2 (MEN, n.d.), the test aimed to evaluate learners' attainment of this target (MEN, n.d.). Also, while the French and German tests focused on reading and writing only, the TDE team decided to extend the English test with a listening section, to more comprehensively reflect the scope of the curriculum and with the aim of generating positive washback.[3]

The TDE team set up a test development cycle that adhered to conventional assessment production processes (see Green, 2014) and guidelines of good practice in language testing (see EALTA, 2006; ILTA, 2000, 2007). An important starting point was the development of test specifications, in which the team defined the test construct in relation to the curricular aims and the basic language proficiency target of CEFR level A2. The TDE team, with support from the consultants, then operationalised the specifications through draft items and tasks, which were moderated, piloted, and analysed prior to the test's administration. As described in Brunfaut and Harding (2018), a key step in the test cycle was to establish the alignment of the test with the CEFR through the process of standard setting.

[2] Note that this test was implemented as only one among a range of classroom-based assessments which contributed towards a learner's end-of-year mark. The *Épreuve Commune for English* is therefore a low-stakes test for the individual learner (https://portal.education.lu/epreuvescommunes/Home.aspx).

[3] The testing of speaking remained part of teacher-based classroom assessment.

The data for this article is drawn from the standard setting sessions of the reading and listening sections of the *Épreuve Commune for English.*[4]

Aim

The standard setting meetings for the *Épreuve Commune for English* provided a unique opportunity to gather insights into the macro-level, systemic factors that fed into decision-making. Our aim in this study, then, was to locate such instances of macro-level/systemic influences in the discourse of standard setting participants. In so doing, we aimed to map out different types of influencing factors, and to draw implications for theorising standard setting more generally when international frameworks are applied in local settings.

Methods

The data used in this study are audio recordings of naturalistic discourse produced during standard setting sessions for the *Épreuve Commune for English*, held in Luxembourg in 2014. The following sub-sections provide details on the participants and the nature of the standard setting approach, the data collection procedure, and the methods of analysis.

Participants

The standard setting sessions were led by the two language testing consultants from Lancaster University (the authors of the study), who had been working with the TDE team since 2011, and who are referred to as the 'moderators' in this article. Both have several years of experience in standard setting – as moderators and as judges – for a range of language

[4] Note that benchmark scripts for writing, which is marked by teachers using an analytic rating scale developed by the TDE team, were determined on a different occasion.

exams. The standard setting panel itself consisted of ten local English language teachers who participated as judges. Seven of these had prior experience as judges in standard setting in the context of the CEFR. The group represented teachers from a range of secondary schools and school streams within Luxembourg.

Standard setting approach

The standard setting sessions were held over two days at the Ministry of Education in Luxembourg City. The focus of the sessions was the listening and reading components of the *Épreuve Commune*. Prior to the standard setting session, two booklets including different sets of potential test tasks and items had been trialled with a representative sample of the target population (N = 350). Details of the trial sample are provided in Table 1.

Table 1. Trial details

Booklet 1	**Booklet 2**
Total sample: n = 179	Total sample: n = 171
4ème ES classique: n = 58	4ème ES classique: n = 57
5ème ES moderne: n = 26	5ème ES moderne: n = 26
10ème EST Régime technique: n = 76	10ème EST Régime technique: n = 72
10ème EST Régime formation technicien: n = 19	10ème EST Régime formation technicien: n = 16

A final version of the listening and reading sections of the test, each consisting of four tasks and 30 items, was compiled on the basis of task quality as determined by the trial. Statistics for items which were included in the final set of tasks in the administered test were extracted for the purposes of providing item facility values in the standard setting sessions.

We used a Basket Method – one of the standard setting approaches described in the *Manual* for relating language exams to the CEFR (Council of Europe, 2009) – which is an

intuitive and simple, but frequently used standard setting technique in language testing (see e.g. Alderson, 2005). Participants are required to perform the test tasks themselves, and then go back through each task to determine what CEFR level each particular item is assessing. First, in accordance with the *Manual* for relating language exams to the CEFR, a number of CEFR familiarisation activities were conducted with the participants (i.e. the standard setting judges) at the start of the session. Next, participants were provided with the test booklets (as students would see them), a copy of the relevant CEFR scales, and a response sheet to record the levels they assigned to each item. Participants were also assigned an anonymised ID code. Participants then completed the tasks and provided their judgements on each item.

After following this procedure for each task (roughly 7-10 items), responses were collated and entered into a spreadsheet which showed all participants' IDs, their judgements on each item, and the item facility values drawn from the test trial. This spreadsheet was displayed, and a plenary discussion held in which points of difference between the judges were discussed, and participants had the opportunity to defend or critique assigned CEFR levels. Following the plenary discussion, participants were given the chance to change their responses if they wished (though there was no pressure to do this), before submitting their final responses to the moderators. This procedure was repeated for each task across the listening and reading tests.

Data collection

As described above, the standard setting process involved multiple rounds of judgement interspersed with plenary discussion of each reading and listening task. At the end of the standard setting, a general discussion was also held in which participants were asked to reflect on the standard setting process as a whole. Both the test task-related discussions and

the general discussion were audio-recorded. These recordings were then transcribed by a research assistant.

Data analysis

The discourse produced during the standard setting discussions and the follow-up group reflection on the standard setting process formed the focus of the data analysis. Given the exploratory aim of the study, no *a priori* analytic framework was adopted, but an inductive approach was followed to let themes emerge from the data in a bottom-up manner following a 'thematic analysis' approach (Braun & Clarke, 2006). We each independently read through the transcripts, identifying parts of the discourse that were judged to be salient to the aim of the analyses, i.e. to explore participants' challenges in applying the CEFR, tracing the influence of the broader exam reform context, and the particular nature of languages education in the Luxembourg school system. Throughout this process we adhered to principles of conducting thematic analysis in such a way as to maximise the 'trustworthiness' of our analysis (see Lincoln and Guba, 1985; Nowell, Norris & White, 2017). We first familiarised ourselves with the data by reading through the transcripts several times, and then each generated an initial set of codes and themes independently. We then held a face-to-face meeting during which we discussed the initial set of themes we had identified independently, and the specific discourse excerpts that illustrated these. This meeting served as a key point of initial researcher triangulation (Nowell, Norris & White, 2017), and the result was a set of detailed notes concerning initial hierarchies of themes, codes and discourse excerpts. The initial exchange, in any case, revealed that both researchers had independently identified similar themes and salient excerpts. We continued to define and refine themes through consensus (essentially, a prolonged research triangulation) in order to ensure credibility of our analysis, and this further discussion allowed us to develop a set of thematic categories.

Finally, we checked 'referential adequacy' (Nowell, Norris & White, 2017) by applying our thematic categories back to the raw data and found a very good fit. The themes which resulted in this analysis are presented in the next section.

Findings

The analysis of the transcriptions revealed discussion around four main themes related to broader systemic influences on the application of the CEFR in the standard setting process. These themes were: (1) the unique multilingual language learning ecology within Luxembourg; (2) the differentiation of ability within the streamed schooling system; (3) the constraints of the national curriculum; and (4) the broader exam reform aims and objectives. Each of these themes is discussed below.

Theme 1: *Luxembourg's multilingual language learning ecology*

As discussed in the introduction, the language ecology of Luxembourg – both within Luxembourgish society and within educational contexts – is unique. This context led to considerations in the standard setting sessions – both within the judgement rounds themselves and in post-hoc discussion – of the nature of language learning in the Luxembourg system. A useful illustration is provided in Excerpt 1, which is drawn from the post-hoc plenary discussion. In this extract, one teacher problematizes the interpretation of item facility values within the basket method on the grounds that high values reflected that students in this trial sample typically scored very high on the test.

[1]

Post-hoc discussion

T5; When we also look into the facility values, especially with the listening for example, … we … of course we should change our judgments. However sometimes there are things maybe our students are just so much better at something. That doesn't therefore then just make

it an A1. And it does make it difficult. There's not always something in the CEFR to point out, but you just say 'that was a bit more difficult'.

Implicit in this comment is the notion that Luxembourg students at lower-secondary level are exceptionally proficient, and that the majority reaches the minimum curricular target level of CEFR A2 with ease. The existence of a ceiling effect is not entirely unusual for an achievement test. However, this broader context of generally high proficiency levels makes standard setting a difficult enterprise as the item statistics come to be considered less trustworthy indicators of how challenging a specific item really is according to the CEFR. This may be understood as an implicit criticism of the basket method and the CEFR: there is no way of adequately indicating an exceptionally strong cohort in the judgement process. However, other standard setting methods would be equally troublesome. For example, when discussing an alternative Angoff method, in which the conceptualisation of a 'minimally competent' test-taker is necessary, the participants in the study revealed further issues regarding the unique nature of language learning in Luxembourg in conceptualising such a figure (see Excerpt [2]).

[2]

Post-hoc discussion

T3; Yeah, we got an interesting discussion started here about picturing this person, minimally
T5; competent
T3; competent A2 student. And if we look at what we found here, it would be somebody who would be able to do forty seven percent of the items. Now we all know that all our students can do that, so
T5; It would have to be someone who is really struggling in 6ème then.
T3; They would still be able to do it.
T5; They would, yeah.
R2; Or the person in the year before.
T5; Yeah, but they don't have
T4; No, they don't have English in the year before.
T6; The 6ème is their first year of English.
T8; Wow.
T6; But they just advance very quickly.

R2; This is just like just by osmosis, not a
T4; No, they have six hours a week.
T8; No, but they have very good teachers.
R2; Of course.
T7; I think what would be interesting […] from a linguistic point of view. I mean they are all at least L4 learners, so they are all expert learners of languages.
Several; <overlapping agreement>
T7; So, they just transfer all the skills that they have acquired.

Excerpt 2 shows that standard setting participants believed even a student deemed to be 'really struggling' in 6ème would be able to reach the A2 level standard. At this level, the CEFR describes language learners as being able, for example, to 'understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment)' (Council of Europe, 2001, p.24). Two points are noteworthy here. First, the participants appeared to agree that their learners acquire English quickly because of their multilingual background, with English being the 'fourth language' ('L4') for many. Therefore, the participants speculated that their learners are able to transfer skills between languages, and can be regarded as 'expert learners of languages' (see T7, Excerpt [2]). Second, this excerpt points to a potential mismatch between the A2 CEFR standard - which is understood to be easily achievable – and the 'real' standard taught in the classroom – with which a student might struggle.

Returning to the first point, the conceptualisation of Luxembourg students as 'language learning experts' was seen to influence judgements in specific standard setting decisions. In Excerpt [3], the judges are discussing a reading item in which comprehending the word 'renovate' is integral to answering the item. The relevant extract from the passage and the associated multiple-choice item is shown in Figure 1.

Figure 1. Reading task extract

Extract from the reading passage:

Farming holidays are going to be a new addition to our offer in 2014. If you want to learn English, enjoy working with animals and want to know more about organic farming, choose a stay in Cornwall, England. This holiday also offers unique beach riding experiences. If you are looking for something more unusual, select our Kenya holiday to visit a coffee plantation and help the local people renovate and improve their old elementary school.	
Associated item:	
During the holiday in Kenya you	A ☐ repair a building. B ☐ teach young children. C ☐ pick coffee beans.

A discussion forms around whether the item might fit a B1 level CEFR descriptor (one level higher than the A2 target level) on the basis that the CEFR A2 descriptors characterise language input at this level as 'simple', 'high-frequency everyday language, and 'containing the highest frequency vocabulary' (Council of Europe, 2001, p.69), which the word 'renovate' appears to exceed. However, T4 argues against a B1 level conclusion as the lexeme 'renovate' would be particularly easy for Luxembourgish students who have cognates in their other languages, French (*rénover*) and German (*renovieren*).

[3]

Reading task

R1; Straightforward and factual – is the B1 descriptor.

T7; I think.

T3; Is it?

R1; Yes, the general B1 descriptor.

T4; But it could also be… Would that be 'shared international vocabulary'? Because you'd have it in German, you'd have it in French: *rénover, renovieren.*

The unique language learning expertise of Luxembourg students thus led to a frequent need for standard setting judges to balance 'canonical' interpretations of the CEFR descriptors with their local knowledge of the abilities of their student population.

Theme 2: *Streamed schooling system*

A second macro-level factor relates to the structure of Luxembourg's secondary schooling system. The system comprises different 'streams', with differences in subjects, focus and depth of subject coverage, speed at which content is covered, etc. Different streams are also designed with alternative pathways beyond secondary school in mind. At the time of the 2014 standard setting, the two main streams were labelled 'Enseignement secondaire' (ES) and 'Enseignement secondaire technique' (EST) (The streams have recently been renamed and year levels renumbered, but without material effect on this study). The ES stream, which comprised classical (Latin) and so-called modern (general) studies, was designed to lead to higher education. The EST stream, on the other hand, was designed to constitute a professional, technical and vocational pathway. The *Épreuve Commune* was taken by ES students in what was called 6ème Moderne and 5ème Classique – at the end of their first year of English classes, and by EST students in 9ème Technique – at the end of their second year of English language instruction.

Streaming considerations were brought to bear in discussions of item difficulty during the standard setting sessions. Excerpt [4], for example, illustrates this influencing factor in deliberations around a listening task. Note that the term 'grammar school' is used here to refer to the ES stream (potentially, specifically the Classique stream).

[4]

Listening task

T6; […] Probably not a problem for the grammar school kids. But the other ones, they'll be listening and think...huh...'What the hell is this?'.

T4; The stupid school book, [T6 name]. No more, no less.

T10; But not in the EST.

This Excerpt is from a discussion on whether two particular items of one of the listening test tasks are at the CEFR A2 (lower) or B1 (higher) level. With reference to a particular chunk of

the task's audio passage, T6 thinks that it may be straightforward to comprehend for the ES stream, but may not be understood by other streams (and therefore T6 seemed to lean towards a CEFR B1 judgement). T4, on the other hand, claims that learners would know what the chunk was about from their school books (so T4 seemed to argue that the level may be lower – CEFR A2). T10, however, expresses support for T6's reasoning that this will be difficult for learners in the EST stream.

Overall, streaming was considered a challenging macro-level factor for two reasons: (1) judges found it difficult to set standards which would apply to all potential test-takers in an equivalent way (as in Excerpt 4 above), and (2) the data used during the standard setting needed to adequately represent the proportion of students across the different pathways.

Theme 3: *National curriculum and textbooks*

Another factor which transpired from the discussions relates to constraints of the *Épreuve Commune* due to the national curriculum. More specifically, the curriculum stipulates the CEFR A2 level as the basic target for English for the relevant years of lower secondary school, and a crucial purpose of the *Épreuve Commune* was to measure achievement against that target. At the same time, the teachers' insights were that many of their students are already more proficient than this curricular target (as illustrated in Theme 1), and so the CEFR A2 level was viewed as a mismatch with the 'real' level of the students. This created a 'dilemma' for judges who felt constrained by the level set in the curriculum, and were aware that the standard they were setting for the test would ultimately paint an unrealistic picture of their students' abilities.

[5]

Post-hoc discussion

R1; But it's always been this dilemma that you've had: on one hand the curriculum says they should reach A2, so you want to show they meet the curriculum requirements .

T1; Yeah.
R1; In reality, you've got a gut feeling they actually – many of them beyond this.
T1; Especially receptive skills.
R1; So it's a different… That struggle between 'do you want to show their actual level' or 'do you want to show they are A2 or above'.

In addition, the content covered by textbooks used in English language teaching in Luxembourg also functioned as a factor influencing standard setting decisions. For example, Excerpt 6 illustrates how T9, on the one hand, acknowledged that the features of one of the listening test tasks matched with descriptors of the CEFR B1 level. For example, the speech in the audio passage corresponded with the B1 criterion of 'clear standard speech' and was delivered at a faster pace than the A2 criterion of 'slowly articulated' (Council of Europe, 2001, p.66). Nevertheless, on the other hand, T9 decided on a lower, CEFR A2 level judgement on the basis of what is covered in the textbooks used in schools and the focus of the item, namely the topic area of countries and languages.

[6]
Listening task
T9; […] The only thing that made me decide against the B1 was basically a fact which is not related to the CEFR at all, but the simple reason that so many of them have actually been studying nationalities and countries in their course books. And so it wouldn't have caused them any major difficulties, after they've heard it once, to write these things down afterwards. That was something that influenced my decision.
T6; Yes, because there is a second listening.
R1; So?
T9; But […] in most of the course books that have been used there is always something on nationalities and languages. There's very often something on that.
T6; And music as well.
T9; And so they might have been able to draw on that knowledge to fill in the correct information here. However, the speed, the delivery…

The nature of the national curriculum and the textbooks used for English language teaching in Luxembourg thus require standard setting judges to interpret item statistics and

item content with reference to the specific test context, and balance this against more narrow interpretations of CEFR level descriptors.

Theme 4: *Exam reform aims/objectives*

Finally, the discourse analysis demonstrated the latent influence of the broader context for standard setting: the teacher-led exam reform project. Excerpts [7] and [8] exemplify some of the reforms' more international ambitions, which at least partly relate to the fact that, given Luxembourg's small size, many learners study abroad at tertiary level. The act of standard setting itself, the rigour with which it is conducted, and the validity of the linking results as such become significant tools of the reform. Thus, the standard setting judges recognised the vital importance of 'doing a good job' at decision-making during the standard setting process to ensure international recognition of the CEFR alignment. The consequences of this were considered crucial, as they may determine key life events of Luxembourg youngsters such as access to tertiary education later, as hinted by T4.

[7]

Post-hoc discussion

T4: […] When our students leave Luxembourg, they need to have some sort of evidence of what they can do which is understandable outside Luxembourg […].

T7; […] It's clear that you have to align to international standards […].

[8]

Post-hoc discussion

T4; I think there are a couple of problems here. One is that it hasn't hit home yet how important levels of English are abroad. I mean, the ministry has had evidence, because students telephone the ministry: 'So I need some sort of reference or something that says what I've done in my A levels'. And they say: 'Well, how has this been done so far…Well, can't we just um…'. And universities do not accept that anymore.

This means that participants were trying to find a balance in their judgements between the localisation of standard setting decisions and their rationales (see Themes 1-2-3) with what would be internationally acceptable CEFR interpretations and decisions during standard setting.

Discussion

In this paper, we have presented a thematic analysis of naturalistic discourse gathered through recordings of standard setting sessions and post-hoc discussions in the context of the Luxembourg *Épreuve Commune for English*. Our aim in the study was to locate and discuss instances of talk where macro-level/systemic factors influenced standard setting decisions, and the process of standard setting generally. Through our analysis, we identified four key factors which appeared to influence the ways in which participants made decisions, and engaged with the standard setting process. These were: (1) the unique multilingual language learning ecology within Luxembourg; (2) the differentiation of ability within the streamed schooling system; (3) the constraints of the national curriculum; and (4) the broader exam reform aims and objectives.

The study has several implications. It is clear that the role of external factors needs to be accounted for, and perhaps included, in theorising standard setting processes. Specifically, it is clear that the role of context in standard setting needs to be articulated unambiguously. The CEFR, as mentioned earlier, is designed to be used across multiple languages and regions and is therefore designed to transcend context (Papageorgiou & Tannenbaum, 2016). Yet, our data reveals that applying the CEFR in standard setting in a specified and unique local educational environment is an activity which is highly context-bound. McNamara (2011) has argued that localisation of a framework like the CEFR is fundamental, since a

dogmatic treatment of the CEFR as a 'common currency' (p.504) comes at a significant cost. On this point, McNamara (2011, p.506) writes that:

> The imposition of a single set of cultural meanings and social and political values for language education, for each setting in which the CEFR is adopted, eviscerates the traditions of language teaching which are incompatible with the CEFR. In cultural and historical terms, learning English is simply not the same for a Singaporean, an Indonesian, a Vietnamese, a French person, a Dutch person, or a Hungarian.
> […]
> The cost of unification is the devaluing of the local interpretation of the goals of education. One issue is that the limiting of the goals and meaning of language learning to functional, communicative objectives ignores the role of language learning in the subjective experience of the learner as an individual with a history, both personal and cultural.

Similarly, in a review of Martyniuk's (2010) edited collection of alignment studies, Deville (2012, p. 313) expressed support for context-bound decision-making during CEFR-alignment activities:

> Certainly allowing experts and practitioners to make local decisions with respect to what evidence best supports their claims of alignment and linkage to the CEFR is positive. Professionals within the local context know how test scores will be used and interpreted, so are in the best position to make decisions with respect to the appropriateness and relevance of the evidence.

However, at the same time, Deville (2012) commented: 'On the other hand, a lack of prescription can leave some practitioners constantly questioning whether they are on the right track with their alignment work' (p.313). Therefore, without further guidance for novice practitioners on how to balance local-contextual demands, there is a risk that standards become applied across contexts in a way that eventually defeats the purpose of having any standards.

In this study, standard setting emerges as a space/activity in which international standards and local educational cultures collide. We therefore need a better way of theorising how to involve local knowledge in the standard setting process without compromising the

validity of procedures (see Papageorgiou & Tannenbaum, 2016). A useful avenue for further research would therefore be to focus on successful practices in combining localisation with international standards such as the CEFR. Such work would need to be conducted across multiple languages and educational contexts. Following the recommendations of Manias and McNamara (2016), further qualitative, discourse-analytic work is also required for routine standard-setting studies to reveal the bases for concerns, decisions and rationales, as validation of procedures must extend beyond quantitative measures of judge and cut-score reliability.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Allais, S. (2014). *Selling out education: National Qualification Frameworks and the neglect of knowledge*. Rotterdam/Boston/Taipei: Sense Publishers.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101.

Brunfaut, T., & Harding, L. (2014). Linking the GEPT listening test to the Common European Framework of Reference. *LTTC-GEPT Research Reports RG-05*. Taiwan: Language Training and Testing Centre. https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG05.pdf

Brunfaut, T., & Harding, L. (2018). Teachers setting the assessment (literacy) agenda: a case study of a teacher-led national test development project in Luxembourg. In D. Xerri, & P. Vella Briffa (Eds.), *Teacher involvement in high stakes language testing* (pp. 155-172). Springer.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Council of Europe (2001). *The Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A manual.* Retrieved January 2019, from https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Deville, C. (2012). Book review: Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual. *Language Testing*, *29*(2), 312-314.

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly*, *15*(1), 3-15.

EALTA (2006). *EALTA Guidelines for good practice in language testing and assessment*. European Association for Language Testing and Assessment. Retrieved October 2011 from http://www.ealta.eu.org/guidelines.htm

European Commission/EACEA/Eurydice (2017). *Key data on teaching languages at school in Europe – 2017 edition*. Eurydice report. Luxembourg: Publications Office of the European Union. Retrieved January 2019 from https://eacea.ec.europa.eu/national-

policies/eurydice/content/key-data-teaching-languages-school-europe-%E2%80%93-2017-edition_en

Figueras, N., & Noijons, J. (2009). *Linking to the CEFR levels: Research perspectives*. Arnhem: CITO and EALTA.

Geyer, F. (2009). *The educational system in Luxembourg*. CEPS Special Report. Centre for European Policy Studies. Retrieved January 2019, from http://aei.pitt.edu/14574/1/Includ-ed_FG_on_Ed_System_in_Luxembourg.pdf

ILTA (2000). *ILTA Code of ethics*. International Language Testing Association. Retrieved October 2011 from http://www.iltaonline.com/index.php/en/resources/ilta-code-of-ethics

ILTA (2007). *ILTA Guidelines for practice*. International Language Testing Association. Retrieved October 2011 from http://www.iltaonline.com/index.php/en/resources/ilta-guidelines-for-practice

Johnson, S., & Wolf, A. (2009). Qualifications and mobility in a globalising world: why equivalence matters. *Assessment in Education: Principles, Policy & Practice, 16*(1), 3-11.

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing, 13*(1), 32-49.

Lincoln, Y., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Luxembourg.lu (n.d.). *Emigration and immigration*. Le Gourvernement du Grande-Duché du Luxembourg. Retrievend January 2019 from http://luxembourg.public.lu/en/le-grand-duche-se-presente/population/emigration-immigration/index.html

Manias, E., & McNamara, T. (2016). Standard setting in specific-purpose language testing: What can a qualitative study add? *Language Testing, 33*(2), 235-249.

Martyniuk, W. (2011). *Aligning tests with the CEFR. Reflections on using the Council of Europe's Draft Manual*. Cambridge: Cambridge University Press.

McNamara, T. (2011). Managing Learning: Authority and language assessment. *Language Teaching, 44*(4), 500-515.

Men.lu (n.d.). *Langues à l'école luxembourgeoise*. Ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse. Retrieved January 2019 from http://www.men.public.lu/fr/themes-transversaux/langues-ecole-luxembourgeoise/index.html

MEN (n.d.). *Syllabi: Enseignement secondaire technique – Cycle inférieur, 8e et 9e théorique, Anglais / Division inférieure 6e moderne et 5e classique, Anglais*. Luxembourg: Ministère de l'Éducation nationale, et de la Formation professionnelle.

MEN (2010). *Dossier de presse: La réforme des classes inférieures de l'enseignement secondaire at secondaire technique*. Luxembourg: Ministère de l'Éducation nationale, et de

la Formation professionnelle. Retrieved January 2019 from http://www.men.public.lu/fr/actualites/publications/themes-transversaux/dossiers-presse/2010-2011/100929-reforme-classes-inferieures/index.html

MEN (2011a). *Dossier de presse: La réforme des classes superérieures de l'enseignement secondaire at secondaire technique*. Luxembourg: Ministère de l'Éducation nationale, et de la Formation professionnelle. Retrieved January 2019 from http://www.men.public.lu/fr/actualites/publications/themes-transversaux/dossiers-presse/2010-2011/110512-classes-sup/index.html

MEN (2011b). *Dossier de presse: Réforme du lycée. Proposition de texte d'une loi sur l'enseignement secondaire*. Luxembourg: Ministère de l'Éducation nationale, et de la Formation professionnelle. Retrieved January 2019 from http://www.men.public.lu/fr/actualites/publications/themes-transversaux/dossiers-presse/2011-2012/111205-dossier-presse-texte-loi/index.html

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1-13.

Papageorgiou, S. (2009). *Setting performance standards in Europe: The judges' contribution to relating language examinations to the common European framework of reference*. Frankfurt am Main, Germany: Peter Lang.

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly, 13*(2), 109-123.

Spöttl, C., Kremmel, B., Holzknecht, F., & Alderson, J. C. (2016). Evaluating the achievements and challenges in reforming a national language exam: The reform team's perspective. *Papers in Language Testing and Assessment, 5*(1), 1-22.

Tannenbaum, R. J., & Wylie, E. C. (2008). Linking English-language test scores onto the common European framework of reference: An application of standard-setting methodology. *ETS Research Report Series, 2008*(1), i-75.

Visitluxembourg.com (n.d.). *Key facts*. Retrieved January 2019 from http://www.visitluxembourg.com/en/travelguide/key-facts

Young, M., & Allais, S. M. (2013). *Implementing national qualifications frameworks across five continents*. London/NY: Routledge.