CEFR Journal

Research and Practice

Japan Association for Language Teaching (JALT)

CEFR & Language Portfolio SIG (CEFR & LP SIG)

Volume 7 (March 2025)

ISSN 2434-849X

Title: CEFR Journal—Research and Practice

Type: Online Journal
URL: https://cefrjapan.net/publications/journal
Contact: journal@cefrjapan.net

ISSN: 2434-849X
DOI https://doi.org/10.37546/JALTSIG.CEFR

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Table of Contents

Mission statement

The CEFR Journal is an online, open-access, peer-to-peer journal for practitioners and researchers. Our editorial advisory board comprises stakeholders on a wide range of levels and from around the world. One aim of our journal is to create an open space for exchanging ideas on classroom practice and implementation related to the CEFR and/or other language frameworks, as well as sharing research findings and results on learning, teaching, and assessment-related topics. We are committed to a strong bottom-up approach and the free exchange of ideas. A journal by the people on the ground for the people on the ground with a strong commitment to extensive research and academic rigor. Learning and teaching languages in the 21st century, accommodating the 21st century learner and teacher. All contributions have undergone multiple double-blind peer reviews. We encourage you to submit your texts and volunteer yourself for reviewing. Thanks a million.

Aims, goals, and purposes

Our aim is to take a fresh look at the CEFR and other language frameworks from both a practitioner's **and** a researcher's perspective. We want the journal to be a platform for all to share best practice examples and ideas, as well as research. It should be globally accessible to the wider interested public, which is why we opted for an open online journal format.

The impact of the CEFR and now the CEFR Companion Volume (CEFR/CV) has been growing to previously wholly unforeseeable levels. Especially in Asia, there are several large-scale cases of adoption and adaptation of the CEFR to the needs and requirements on the ground. Such contexts often focus majorly on English language learning and teaching. However, there are other language frameworks, such as the ACTFL and the Canadian benchmarks, and the Chinese Standard of English (CSE). On the one hand there is a growing need for best practice examples in the form of case studies, and on the other hand practitioners are increasingly wanting to exchange their experiences and know-how. Our goal is to close the gap between research and practice in foreign language education related to the CEFR, CEFR/CV, and other language frameworks. Together, we hope to help address the challenges of 21st century foreign language learning and teaching on a global stage. In Europe, many take the CEFR and its implementation for granted, and not everyone reflects on its potential uses and benefits. Others are asking for case studies showing the effectiveness of the CEFR and the reality of its usage in everyday classroom teaching. In particular, large-scale implementation studies simply do not exist. Even in Europe, there is a center and a periphery of readiness for CEFR implementation. It is difficult to bring together the huge number of ongoing projects from the Council of Europe (CoE), the European Centre for Modern Languages (ECML), and the EU aiming to aid the implementation of the CEFR. This results in a perceived absence in the substance of research and direction. Outside Europe, the CEFR has been met with very different reactions and speeds of adaptation and implementation. Over the last few years, especially in Asia, the demand by teachers for reliable (case) studies has been growing.

For more than a decade, the people behind this journal – the Japan Association for Language Teaching (JALT) CEFR & Language Portfolio special interest group (CEFR & LP SIG) – have been working on a number of collaborative research projects, yielding several books and textbooks, as well as numerous newsletters. This is a not-for-profit initiative; there are no institutional ties or restraints in place. The journal aims to cooperate internationally with other individuals and/or peer groups of practitioners/researchers with similar interests. We intend to create an encouraging environment for professional, standard-oriented practice and state-of-the-art foreign language teaching and research, adapted to a variety of contexts.

Editorial advisory board

Journal editorial team

Editing and proofreading team

Layout

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Editorial

Fergus O'Dwyer

The underlying philosophy of open scholarship—freely sharing knowledge in a collaborative manner—is central to the CEFR, and found in the two contributions that bookend this seventh issue of the journal. The steering group of the freely available *Aligning Language Education with the CEFR: a Handbook* report on an event in Barcelona in 2024 that advanced a wide range of policy and pedagogical issues, with results feeding forward to objectives for future development. This includes an upcoming special guest-edited issue of the *CEFR Journal* on alignment issues, as well as involving international networks of language teacher educators and language teachers. The latter is in a similar vein to the first article by Brian North who discusses the formation of the *Action-oriented, Plurilingual and Intercultural Education* (API) Forum. This new organization aims to facilitate an inclusive and democratic language education that promotes an Action-oriented Approach and plurilingualism. One critical task is the networking of initiatives and projects concerning the key concepts of the CEFR/CV. Another stated aim of the API forum—supporting the development of teacher competences—is threaded through two of the other articles in this issue.

Osidak, Vogt and Natsiuk examine the Ukrainian tertiary-level context, highlighting the importance of the CEFR/CV descriptors in facilitating a shift towards plurilingual assessment. The article builds toward a collaborative approach to knowledge construction that empowers teachers allowing them to be agents for educational change. Jana Bérešová finds that active involvement in rating written work by student teachers in Slovakia leads to an awareness of the many possibilities of applying the CEFR in various contexts, ultimately positively impacting the development of classroom materials and the learning process in general.

The remaining two articles deal with the further development of the CEFR, with Aziza Zaher suggesting the need for a tailored framework that considers the specific features of the Arabic language. Abdulhaleem and Harsch verify the use of the CEFR as a criterion-referenced tool for gaining a broad understanding of proficiency levels, even if participants possess limited familiarity with the CEFR scales.

The above are examples of the broad range of issues which can be addressed in a focused manner by contributing to the journal: we welcome future submissions starting with an upcoming call for abstracts (details to follow on the journal website). Furthermore, we can expect that the emerging connectedness brought about by two initiatives outlined in the first paragraph above will be further developed in the API forum hybrid conference to be held in Rome on 19-20 June 2025. Onward and forward!

—Dublin (Ireland), March 2025

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Action-oriented, Plurilingual and Intercultural Education: A new association—API Forum

Brian North, CEFR co-author; President API Forum

https://doi.org/10.37546/JALTSIG.CEFR7-1

Despite developments over the past 20 to 30 years in theories that inform language education, the predominant pedagogical approach in English Language Teaching (ELT) has not changed radically since the introduction of the coursebook-dominated 'mature version' of communicative language teaching (CLT) in the late 1980s and early 1990s. In the field, the CEFR appears to have been perceived mainly as a set of proficiency levels and descriptors (common reference points) to guide assessments, provide standards and align planning, teaching and assessment. In great part, implementation of the CEFR overlooks the action-oriented, plurilingual approach to language education advocated in the 2001 original, and even more so in the 2020 CEFR Companion Volume. After outlining why that might be the case and briefly summarising the key aspects of the CEFR pedagogic vision – action-orientation, the social agent, mediation and plurilingualism – this article introduces a new international association for language educators and researchers, API Forum, dedicated to promoting, implementing, researching and further developing this vision.

Key words: Innovation; CEFR; Action-orientation; Plurilingualism; Mediation; Professional Associations

1. Impediments to moving forward in language education

Over the past 20 to 30 years, there has been a somewhat static situation in the evolution of second/foreign language teaching practices, when compared to the radical advances made in the 1970s and 1980s that led to the communicative approach, which then became 'in vogue' during the 1990s (Pearce 2024). It is true that pedagogy for many languages[1] went through a 'catch up' transformation in the early 2000s, largely due to the influence of the CEFR, but since the spread of the communicative approach in the 1990s, rather little has changed in classroom practices, particularly in English Language Teaching (ELT) (Piccardo 2024). This is partly because, following the sweeping success of *Headway* (Soars & Soars 1987) the first 'mega coursebook' (Keddle 2004), which reverted to a grammatical progression rather than authentic materials (Keddle 2004), most publishers quickly followed suit (Jordan & Gray 2019; Thornbury 2016). The result was that, as Jordan and Gray suggest, already by 2008 "communicative language teaching (CLT) had been so completely replaced by coursebooks that CLT was now 'part of history'" (2019: 438). It is true that some ELT publishers have since become more adventurous – perhaps starting with the *Speakout* series (Clare & Wilson 2011), but as Jordan and Gray summarize "it is fair to say, following Akbari (2008), that the methodology these coursebooks implement is the current model for ELT worldwide" (2019: 440). The effect has been to reinforce a *deficiency perspective* focused on mistakes rather than the

1. From the early 1980s both English and German had well-established applied linguistics cultures, empirical second language acquisition research, analysis of classroom discourse, teacher training institutes, curriculum approaches, and course book expertise – including, for German, a well-developed course book evaluation methodology. This was far less the case for Romance languages before the CEFR and presumably this was the case for other languages as well.

proficiency perspective promoted by the CEFR: what you 'can do.' Such conservatism is further reinforced by the dominant culture of standardised tests, which still follow Lado's (1961) model of separating the 'four skills' with decontextualized test tasks and sometimes even separate language papers. As Akbari pointed out 15 years ago, "the profession has not yet been able to counter the destructive effects of standardized tests" (2008: 649). Since then, things have not greatly improved. The CEFR Expert Group give the PISA 2025 Framework as an example, since it "eschews interaction and continues to test the four skills (listening, reading, spoken production, and writing) in complete isolation, using picture-based item types to do so that would not have been out of place in the 1970s" (2023: 28).

Another factor impeding a move forward in language education is the way that, despite the CEFR's provision of a common metalanguage, in teaching, teacher education, and testing, languages still tend to be kept in complete isolation from one other. This is partly due to the influence of the monocultural/monolinguistic institutes and examination bodies associated with each language (e.g., British Council, Cambridge, France-Éducation-Internationale, etc.), which leads to the maintenance of separate pedagogic cultures for different languages. The separation of languages at a school level impedes the leveraging of new learning in relation to language(s) known, hindering the development of metalinguistic awareness on the part of both teachers and learners. The form taken by teacher education also exacerbates the isolation problem since it generally takes place separately for different languages. In addition, pre-service education institutes often lack the power to select the teachers that host trainees for teaching practice. This fact – plus the power of textbooks and tests and the way new teachers are influenced by the way they themselves were taught languages – contributes to the tendency for new teachers to fall back on outdated methodologies. Then again, continuous professional development is not as common as it could and should be and often involves one-off events like conferences rather than opportunities to work with other teachers on new ideas, try them out in class, and return to discuss experiences.

2. The CEFR project

It is important to raise awareness of these challenges and share possible solutions to them. In this respect, it is worth remembering that the CEFR project[2] was always intended to help in such a process. The CEFR 2001 had two aims: (a) to provide a common metalanguage (descriptive scheme) and common references points (levels) to increase transparency and coherence within and between institutions/educational systems; and (b) to stimulate reflection and reform in language education, which is why each chapter ended with questions for the reader to consider (generally referred to as 'reflection boxes'). At the intergovernmental Forum held in 2007 to take stock of implementation of the CEFR, member states made clear that, while they recognised that it was important to respect the integrity of the scheme and levels in respect of the first aim, they were *far* more interested in the second (Goullier 2007a; 2007b). Unfortunately, as many have pointed out, in practice success with the first aim largely eclipsed the second[3] (see Byram & Parmenter 2012; Coste 2007; North 2014; Piccardo 2020; Piccardo & North 2019; Savski 2019, 2020). There are of course CEFR-focused Special Interest Groups (SIGs) in some national teacher associations, which concern themselves with educational aspects of the CEFR (e.g., the JALT CEFR LP SIG that is the parent of this journal), but these are the exception.

The Common European Framework of Reference for Languages: Learning, teaching, assessment –

2. By CEFR project I refer to the ongoing development of ideas in the CEFR, which includes the CEFR 2001, the CEFR Companion Volume and the CEFR website (www.coe.int/lang-cefr). In Chapter 1 of the CEFR 2001 it was made clear that the Framework was seen as an open-ended project "open: capable of further extension and refinement; dynamic: in continuous evolution in response to experience in its use." (Council of Europe 2001: 8)
3. This process can be seen to be continuing with, for example, the publication of the recently published handbook for Aligning Language Education with the CEFR (Figueras et al. 2022), which is essentially an update of the manual for relating assessments to the CEFR (Council of Europe 2009), and which interprets 'aligning language education' purely in terms of levels, without considering educational aspects.

Companion Volume (CEFR/CV: Council of Europe 2020) was produced in an attempt to address this problem by spelling out and further developing the key concepts of the CEFR vision – such as the action-oriented approach, the learner as a social agent, mediation and plurilingualism – which can help to address the problems mentioned above. These key concepts, which are explained and illustrated on the CEFR website (www.coe.int/lang-cefr), are very briefly outlined below.

Action-orientation: A few pioneers in French as a foreign language (Bourguignon 2006, 2010; Piccardo 2005; Puren 2004, 2009; Richer 2009; Rosen 2009) as well as van Lier (2007) saw the implications of an action-oriented / action-based approach in the early 2000s but it was not until more recently that the approach was explained (Piccardo 2014) and theorised (Piccardo & North 2019) or that scenario-based, action-oriented classroom materials have begun to appear (e.g., Collins & Hunter 2013, 2014; Hunter et al. 2019; Lebrec et al. 2024; Piccardo et al. 2022a), often in the context of teaching adult immigrants (e.g., Durham Immigration Portal 2016; Hunter et al. 2017; Piccardo & Hunter 2017; Schleiss & Hagenow-Caprez 2017). The action-oriented approach aligns completely with an ecological approach (van Lier 2000, 2004, 2010), complexity theories (Larsen-Freeman 1997, 2011, 2017) and the socio-cultural theory (Lantolf 2000, 2011; Lantolf & Poehner 2014). Unlike the communicative approach, it foregrounds learner agency in situated, collaborative learning (Webb 2009). Learning needs to be experiential, rooted in *dynamic learning situations* (Masciotra & Morel 2011).

The social agent: The action-oriented approach sees the learner as a social agent: acting collaboratively and responsibly with others in a specific context to complete tasks that build up to the production of some kind of performance or artifact, and – through a process of drafting/redrafting/rehearsing, with scaffolding from the teacher – mobilising and extending all their linguistic resources and general competences. In the socio-cognitive theory of agency (Bandura 2001, 2018) agency is developed through forethought (including having some kind of plan); self-reactiveness (self-regulation) and self-reflectiveness. The effect of having such agency is to promote engagement and self-efficacy (the belief in success based on experience of success).

Mediation: This concept was introduced in the CEFR 2001 in a limited form and has been incorporated in assessment in Germany (Katelhön & Marečková 2022; Katelhön & Nied Curcio 2013; Kolb 2016; Reimann & Rössler 2013), Greece (Dendrinos 2013, 2022, 2024; Stathopoulou 2015), Austria (Piribauer et al. 2015; Steinhuber 2022) and more recently Spain (Berceruelo Pino et al. 2024; Sànchez Cuadrado 2022). Central to the sociocultural theory, mediation was theorised and further developed in a 2014-2020 Council of Europe project (North & Piccardo 2016) and is a core feature of the CEFR/CV. It encompasses both intralinguistic mediation and cross-linguistic mediation and can be seen as linguistic, textual, social, cultural and pedagogic. The CEFR/CV provides a wealth of mediation descriptors articulated into different aspects of mediating a text, mediating concepts, and mediating communication. The descriptors have stimulated considerable innovation (see North et al. 2022; Stathopoulou et al. 2023) including articles in this journal (e.g., Berceruelo Pino et al. 2024; Jiménez Naranjo et al. 2024; Lankina & Pect 2020; Liontou & Braidwood 2021; Pavlovskaya & Lankina 2019; Perevertkina et al. 2020; Stathopoulou 2020).

Plurilingualism: Plurilingualism posits a single, holistic linguistic repertoire encompassing all the languages, varieties and registers encountered in one's life trajectory (CEFR 2001; Beacco 2005; Piccardo 2018). The concept aligns with complexity theories (Piccardo 2017; Larsen-Freeman & Todeva 2022) and with developing creativity and interculturality (Furlong 2009; Piccardo 2017, 2019). Plurilingualism as an educational philosophy implies the valuing of home languages in the class, a coherent approach to language across the curriculum and the development of openness to new languages and cultures (Beacco et al. 2016). However, although it was well presented in the CEFR 2001 (unlike the action-oriented

approach, social agent or mediation), plurilingualism took some time to take off despite the fact that, as John Trim, the father of the CEFR project, stressed in 2007:

> "**Both at individual and societal levels the concept** [of plurilingualism] **is dynamic**, since the components from the experience of different language and cultures interpenetrate and interact, forming something new, enriched and in continual development. This approach meets better the realities of globalisation than various forms of purism which regard each language and culture as a separate entity, to be preserved and protected against the threat offered by alien forces. Most users of the CEFR have applied it only to a single language but its descriptive apparatus for communicative action and competences, together with the 'can-do' descriptors of levels of competence, are a good basis for **a plurilinguistic approach to language across the curriculum, which awaits development.**" (Trim 2007, emphasis added)

Although there is a substantial literature on plurilingualism, at least in French (see Moore 2019; Moore & Gajo 2009), apart from *Eveil aux langues* (Candelier 2003), CONBAT (Bernaus et al. 2011) and CARAP/FREPA (Candelier et al.,2012), it is not until more recently that one has seen a 'language across the curriculum' approach and the development of plurilingual classroom pedagogies (see Beacco et al. 2016; Brinkmann et al. 2022; Camilleri Grima 2021; Choi J. & Ollerhead 2018; Corcoll López & González-Davies 2016; Daryai-Hansen et al. 2015; Galante et al. 2019, 2022; Jentges et al. 2022; Lau & Van Viegen 2020; Lory & Valois 2021; Piccardo 2013; Piccardo & Langé 2023; Piccardo et al. 2022b: Part V; Prasad 2014, 2015).

3. A New Association: API Forum

In order to provide an international network for those working with, conducting research on and further developing the core concepts described above and, in general, contributing to a conceptual shift towards action-oriented, plurilingual and intercultural education, the association API[4] Forum was formally founded at its first General Meeting on 3 February 2025. The Forum has grown from a network of language professionals who have been working with the CEFR/CV, which aims to redress the balance in the exploitation of the CEFR, emphasising its conceptual vision rather than just its Common Reference Levels (North 2007a, 2007b), as described above. Following an online conference in December 2020 at which 800 people participated, the CEFR Expert Group organised a 2021-2023 series of monthly online workshops explaining the key aspects of the CEFR vision – such as transparency and coherence; the learner as social agent; action-orientation; mediation; plurilingualism – and the way in which the CEFR/CV further develops these. The materials for all these workshops, as well as other useful videos, key documents, articles, and training materials on these CEFR key concepts, plus ideas for implementing them, are all available on the CEFR website (www.coe.int/lang-cefr).

On 15 June 2023, the CEFR Group held a hybrid Reflection Day in Strasbourg, "*The CEFR Companion Volume: Enhancing engagement in language education*," which was attended by 40 experts in language policy, curriculum design, and teacher education from 20 countries, and at which it was decided to form a network. At the same time, the Group was preparing a *Guide to Action-oriented, Plurilingual and Intercultural Education* (CEFR Expert Group 2023), which is now available online on the CEFR website in English and French. The network met again online on 4 June 2024 and decided to form an association, which has been since formalised as API Forum, with its seat in Lausanne, Switzerland. As mentioned, the first General Meeting was held recently, at which the Articles of Association and Founding Committee were confirmed. The first API Conference "*Action-oriented Plurilingual and Intercultural Education: A needed shift in Language Education*" will take place in hybrid form at La Sapienza Università di Roma on 19-20 June 2025. The full programme is available on the conference website.[5]

4. An api in Italian is a bee. We see ourselves as 'pollinators'.
5. The link to the conference website is: https://sites.google.com/view/api-conference-2025

4. Aims of API Forum

API Forum aims to publicize and promote the latest research in the field of language education oriented to action, plurilingualism and interculturality, as well as innovative projects and relevant resources developed in different contexts, in order to support policy development and implementation. Above all, the Forum aims to provide an international, collaborative space – physically and virtually – for members in different contexts to discuss aspects of action-oriented, plurilingual and intercultural education, reflect on the challenges and opportunities that they present at the classroom, institutional and policy-making levels, and share practical examples of implementation and other resources. In the process the Forum will draw attention to and collaborate in relevant academic research, as well as development projects. In the longer term we aim to create, share and disseminate resources, including exemplar action-oriented scenarios and tasks, as well as to support the further development of teacher competences in pre- and in-service teacher education, particularly competences in relation to action-orientation, plurilingual pedagogies and intercultural approaches.

In the context of supporting education for democracy, respect for human rights, inclusive education, learner agency and the valuing of learners' plurilingual and pluricultural repertoires in line with the Council of Europe's Recommendation CM/Rec (2022) *The Importance of Plurilingual and Intercultural Education for Democratic Culture,*[6] API Forum aims to promote activities such as the following:

- the networking of initiatives and projects concerning the key concepts of the CEFR/CV;
- the integration of the CEFR/CV into the professional development of teachers of all subjects in initial and in-service training;
- the development of the role of mediation as a facilitator for learning across the curriculum, encouraging research and case studies;
- the development of the knowledge of and uses of technologies, multimodality as well as Artificial Intelligence for action-oriented, plurilingual and intercultural education;
- the creation of communities of practice for people working according to the principles of action-oriented, plurilingual and intercultural education at local and regional level; and:
- collaboration between educational institutions and sectors from the local to the global level.

The fact of the matter is that in language education, at an international level, there has, up until now, been no association that brings together policy makers, researchers, curriculum developers, teacher educators and teachers of different languages, let alone one dedicated to innovation. There is AILA (*Association internationale de linguistique appliquée*), but it is academic and expensive; there is FIPLV (*Fédération internationale de professeurs de languages vivantes*), but it is an association of associations; there are national and international language teachers' associations, but usually for just one language. Whereas European language testers have ALTE (*Association of Language Testers in Europe*) and EALTA (European Association for Language Testing and Assessment), language education has only EAQUALS (formerly *European Association for Quality Language Services*, and now *Evaluation and Accreditation of Quality Language Services*), but its members are largely private sector language schools and the focus is on quality management, not innovation.

API Forum aims to fill this gap, becoming a space to exchange the latest news, research and ideas as well as to be a catalyst that broadens the scope of language education, giving learners agency in their learning process through an action-oriented approach and promoting plurilingualism and interculturality. In a world increasingly dominated by the overwhelming spread of English, to the extent that the learning of other additional languages is seriously endangered, and in which nativist, xenophobic, neo-fascist ideologies are gaining ground, the promotion of an inclusive rather than instrumental approach to language education is more important than ever.

6. The link to the Recommendation is https://rm.coe.int/1680a967b4

New members are welcome and applications can be made to the Secretary (rmargonis*[admark]*hotmail.com) or President (brianjohnnorth3*[admark]*gmail.com).

5. References

Akbari, Ramin. 2008. Postmethod discourse and practice. *TESOL Quarterly* 42/4. 641–52. https://doi.org/10.1002/j.1545-7249.2008.tb00152.x

Bandura, Albert. 2001. Social cognitive theory: An agentic perspective. *Annual Review of Psychology 52.* 1-26. https://doi.org/10.1111/1467-839X.00024

Bandura, Albert. 2018. Toward a psychology of human agency: Pathways and reflections. *Perspectives on Psychological Science 13*(2). 130-136. https://doi.org/10.1177/1745691617699280.

Beacco, Jean-Claude. 2005. *Languages and language repertoires: Plurilingualism as a way of life in Europe* - A reference study. Strasbourg: Council of Europe. https://rm.coe.int/16802fc1ba (accessed 27 December 2024)

Beacco, Jean-Claude, Byram, Michael, Cavalli, Marisa, Coste, Daniele., Egli Cuenat, M., Goullier, F., & Panthier, J. 2016. *Guide for the development and implementation of curricula for plurilingual and intercultural education*. Strasbourg: Council of Europe Publishing. https://rm.coe.int/09000016806ae621 (accessed 27 December 2024)

Berceruelo Pino, M. T., Diaz Cobo, A., Deseada López Fernandez, M., & Romero Rodriguez, A. 2024. Developing CEFR-based mediation rating scales for standardized exams at state language schools in Andalucia (Spain). *CEFR Journal - Research and Practice 6*. 137-154. https://doi.org/10.37546/JALTSIG.CEFR6-8

Bernaus, Mercè, Áine Furlong, Sofie Jonckheere & Martine Kervran. 2011. *Plurilingualism and pluriculturalism in content-based teaching: A training kit*. Graz: European Centre for Modern Languages. https://www.ecml.at/Portals/1/documents/ECML-resources/2011_11_26_Conbat_fuer_web.pdf?ver=2018-03-20-160408-117 (accessed 27 December 2024)

Bourguignon, Claire. 2006. De l'approche communicative à l' « approche communicactionnelle »: une rupture épistémologique en didactique des langues-cultures [From the communicative approach to the 'communicational approach': An epistemological break in the teaching of languages and cultures]. *Synergies Europe 1*. 58-73. http://gerflint.fr/Base/Europe1/Claire.pdf. (accessed 27 December 2024)

Bourguignon, Claire. 2010. *Pour enseigner les langues avec les CERCL- clés et conseils* [*Teaching languages with the CEFR: Keys points and tips*]. Paris: Delagrave.

Brinkmann, Lisa Marie, Joana Duarte & Sílvia Melo-Pfeifer. 2022. Promoting plurilingualism through linguistic landscapes: A multi-method and multisite study in Germany and the Netherlands. *TESL Canada Journal 38*(2). 88-112. https://doi.org/10.18806/tesl.v38i2.1358

Byram, Michael & Lynne Parmenter. (eds.). 2012. *The Common European Framework of Reference: The globalisation of language policy,* Multilingual Matters, Bristol.

Camilleri Grima, Antoinette. 2021. Implementing plurilingualism in the foreign language classroom. *Lingue antiche e modern 10.* https://doi.org/10.4424/lam102021-9

Camilleri Grima, Antoinette, Michel Candelier, Véronique Castellotti, Jean-François de Pietro, Ildikó Lőrincz, Franz-Joseph Meißner, Muriel Molinié, Artur Noguerol & Anna Schröder-Sura. 2012. *CARAP – FREPA. A framework of reference for pluralistic approaches to languages and cultures: Competences and resources*. Graz: ECML and Strasbourg: Council of Europe. https://www.ecml.at/Resources/ECMLresources/tabid/277/ID/20/language/en-GB/Default.aspx (accessed 27 December 2024)

Candelier, Michel (ed.) 2003. *L'éveil aux langues à l'école primaire. Evlang: Bilan d'une innovation européenne* [Awakening to languages in primary school. Evlang: Assessment of a European innovation]. Bruxelles: De Boeck.

CEFR Expert Group. 2023. *A guide to action-oriented, plurilingual and intercultural education*. Strasbourg: Council of Europe. https://rm.coe.int/1680af9496 (accessed 27 December 2024)

Choi, Julie & Sue Ollerhead (eds.) 2018. *Plurilingualism in teaching and learning: Complexities across contexts.* London & New York: Routledge.

Clare, Antonia & J. Wilson. 2011. *Speakout Elementary Students Book*. Harlow: Pearson Longman.

Collins, Geoff, & Danielle Hunter. 2013. *Scenarios for an action-oriented classroom: Levels A1 & A2. The CEFR in action*. Toronto: RK Publishing.

Collins, Geoff, & Danielle Hunter. 2014. *Scenarios for an action-oriented classroom: Levels B1 & B2. The CEFR in action*. Toronto: RK Publishing.

Corcoll López, Cristina & Maria González-Davies. 2016. Switching codes in the plurilingual classroom. *ELT Journal* 70. 67–77. https://doi.org/10.1093/elt/ccv056

Coste, Daniele. 2007. Contextualising uses of the Common European Framework of Reference for Languages. In Francis Goullier, (ed.), *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities. Intergovernmental Language Policy Forum, Strasbourg, 6-8 February 2007, Report,* 40-49. Strasbourg: Council of Europe. https://rm.coe.int/090000168069b821 (accessed 27 December 2024)

Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* https://rm.coe.int/1680459f97 (accessed 27 December 2024)

Council of Europe 2009. *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*. Strasbourg: Council of Europe. https://rm.coe.int/0900001680667a2d (accessed 27 December 2024)

Council of Europe 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion Volume*. Strasbourg: Council of Europe. https://rm.coe.int/16809ea0d4 (accessed 27 December 2024)

Daryai-Hansen, Petra, Brigitte Gerber, Ildikó Lörincz, Michaela Haller, Olga Ivanova, Hans-Jürgen Krumm & Hans H. Reich. 2015. Pluralistic approaches to languages in the curriculum: The case of French-speaking Switzerland, Spain and Austria. *International Journal of Multilingualism 12*(1). 109–127. https://doi.org/10.1080/14790718.2014.948877

Dendrinos, Bessie. 2013. Testing and teaching mediation. Directions in English language teaching, testing and assessment. Athens: RCeL publications. https://rcel2.enl.uoa.gr/directions/issue1_1f.htm (accessed 27 December 2024)

Dendrinos, Bessie. 2022. A data-driven curriculum with mediation descriptors for plurilingual education. In Little, David & Figueras, Neus (eds.), *Reflecting on the Common European Framework of Reference for Languages and its Companion Volume,* 101-112. Multilingual Matters.

Dendrinos, Bessie. 2024. *Mediation as negotiation of meanings, plurilingualism and language education.* London / New York: Routledge.

Durham Immigration Portal. 2016. *Durham immigration portal: Exploring the region of Durham through task-based learning.* https://www.dce.ca/en/student-services/resources/Synergies-Project/Living-in-English_Discovering-Durham.pdf (accessed 27 December 2024)

Figueras, Neus, David Little, Barry O'Sullivan & British Council. 2022. Aligning language education with the CEFR: A handbook. *British Council, UKALTA, EALTA and ALTE*. https://www.britishcouncil.org/sites/default/files/cefr_alignment_handbook_layout.pdf (accessed 27.12.2024)

Furlong, Áine. 2009. The relation of plurilingualism/culturalism to creativity: A matter of perception. *International Journal of Multilingualism 6*(4). 343-368. https://doi.org/10.1080/14790710903124997

Galante, Angelica, Kerstin Okubo, Christina Cole, Nermine Abd Elkader, Nicola Carozza, Claire Wilkinson, Charles Wotton & Jelena Vasic. 2019. Plurilingualism in higher education: A collaborative initiative for the implementation of plurilingual tasks in an English for academic purposes program at a Canadian university. *TESL Canada Journal 36*(1). 121–133. https://doi.org/10.18806/tesl.v36i1.1305

Galante, Angelica, Maria Chiras, John Wayne N. dela Cruz, & Lana F. Zeaiter 2022. *Plurilingual guide: Implementing critical plurilingual pedagogy in language education*. Plurilingual Lab Publishing. https://escholarship.mcgill.ca/concern/books/0c483q268?locale=en (accessed 27 December 2024)

Goullier, Francis. (ed.). 2007a. *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities. Intergovernmental Language Policy Forum, Strasbourg, 6-8 February 2007, Report*. Strasbourg: Council of Europe. https://rm.coe.int/090000168069b821 (accessed 27 December 2024)

Goullier, Francis. (ed.). 2007b. Impact of the Common European Framework of Reference for Languages and the Council of Europe's work on the new European educational area. In Francis Goullier, (ed.), *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities. Intergovernmental Language Policy Forum, Strasbourg, 6-8 February 2007, Report* (pp. 31-39). Strasbourg: Council of Europe. https://rm.coe.int/090000168069b821 (accessed 27 December 2024)

Hunter, Danielle, Denis Cousineau, Geoff Collins & Gina Hook. 2019. *Action-oriented approach: Handbook*. Ottawa: Canadian Association of Second Language Teachers.

Hunter, Danielle, Enrica Piccardo & Audrey Andrews. 2017. *Synergies: Settlement, integration and language learning / Synergies:* Établissement*, intégration et apprentissage de la langue.* Durham District School Board, Continuing Education. https://www.dce.ca/en/student-services/resources/Synergies-Project/Synergies---English.pdf (accessed 27 December 2024)

Jentges, Sabine, Eva M. Knopp & Paul Sars. 2022. In Venlo gibt es viel Niederländisch, aber auch Englisch und Deutsch: Receptive plurilingualism in the neighbour language. In Brian North, Enrica Piccardo, Tim Goodier, Daniela Fasoglio, Rosanna Margonis-Pasinetti & Bernd Rüschoff (eds.), *Enriching 21st century language education: The CEFR Companion Volume in practice,* 95-108. Strasbourg: Council of Europe Publishing. https://rm.coe.int/1680a68ed0 (accessed 27 December 2024)

Jiménez Naranjo, Macarena, Marga Navarrete & Mazal Oaknín. 2024. The development of mediation strategies when relaying information, both in speech and writing: Task samples for undergraduate students of Spanish as a foreign language. *CEFR Journal - Research and Practice 6*. 104-111. https://doi.org/10.37546/JALTSIG.CEFR6-6

Jordan, Geoff & Humphrey Gray. 2019. We need to talk about coursebooks. *ELT Journal* 73(4). 438–446. https://doi.org/10.1093/elt/ccz038

Katelhön, Peggy, & Marečková, Pavla. (eds.) 2022. *Sprachmittlung im schulischen und universitären Kontext* [Linguistic mediation in the school and university context]*.* Berlin: Frank and Timme.

Katelhön, Peggy & Martina Nied Curcio. 2013. Sprachmittlung - die vernachlässigte Kompetenz in der DaF-Didaktik? Theoretische und sprachpraktische Oberlegungen zur Sprachmittlung in der Germanistik an italienischen Universitäten [Mediation: The neglected competence in the teaching of German as a foreign language? Theoretical and practical considerations concerning mediation in German studies at Italian universities]. *Deutsch als Fremdsprache* 50. 150-158. https://doi.org/10.37307/j.2198-2430.2013.03.04

Keddle, Julia Star. 2004. The CEF and the secondary school syllabus. In Keith Morrow (ed.), *Insights from the Common European Framework,* 43–54. Oxford: Oxford University Press.

Kolb, Elisabeth. 2016. *Sprachmittlung: Studien zur Modellierung einer komplexen Kompetenz [Linguistic mediation: Studies to model a complex competence]*. Münster: Münchener Arbeiten zur Fremdsprachen-Forschung, Waxmann.

Lado, Robert. 1961. *Language testing*. London: Longman.

Lankina, Olga Y., & Pect, Yulia V. 2020. Classroom-based assessment of group discussion: Challenges and opportunities. *CEFR Journal - Research and Practice 3*. 116-125. https://doi.org/10.37546/JALTSIG.CEFR3-8

Lantolf, James P. (ed.) 2000. *Sociocultural theory and second language learning*. Oxford: Oxford University Press.

Lantolf, James P. 2011. The sociocultural approach to second language acquisition: Sociocultural theory, second language acquisition, and artificial L2 development. In D. Atkinson (ed.), *Alternative approaches to second language acquisition,* 24–47. New York, NY: Routledge.

Lantolf, James. P. & Poehner, M. 2014. *Sociocultural theory and the pedagogical imperative in L2 education: Vygotskian praxis and the research/practice divide*. New York: Routledge.

Larsen-Freeman, Diane. 1997. Chaos/complexity science and second language acquisition. *Applied Linguistics 18*(2). 141-165. https://doi.org/10.1093/applin/18.2.141

Larsen-Freeman, Diane. 2011. A complexity theory approach to second language development/ acquisition. In D. Atkinson (ed.) *Alternative approaches to second language acquisition,* 48-72. New York: Routledge.

Larsen-Freeman, Diane. 2017. Complexity theory: The lessons continue. In L. Ortega & Z.H. Han (eds.) *Complexity theory and language development: In celebration of Diane Larsen-Freeman,* 11-50. Amsterdam and Philadelphia: John Benjamins.

Larsen-Freeman, Diane & Elka Todeva. 2022. A Sociocognitive Theory for Plurilingualism: Complex Dynamic Systems Theory. In Enrica Piccardo, Aline Germain-Rutherford & Geoff Lawrence. (eds.), *Routledge Handbook of Plurilingual Language Education,* 209-224. London and New York: Routledge.

Lau, Sunny Man Chu & Saskia van Viegen. (eds.) 2020. *Plurilingual pedagogies: Critical and creative endeavors for equitable language in education*. Springer.

Lebrec, Caroline, Léger, Frenand & Miao Li, with Enrica Piccardo. 2024. *La francophonie en action. B1—Independent User 1.* Toronto: Canadian Scholars.

Liontou, Magdalini & Eva Braidwood. 2021. Mediation in practice in an ESAP course: Versions of the medical English student conference. *CEFR Journal - Research and Practice 3*. 25-42. https://doi.org/10.37546/jaltsig.cefr4-2

Lory, Marie-Paule & Manon Valois. 2021. *Guide d'initiation aux approches plurilingues : Pour une école de langue française* équitable *et inclusive en Ontario [Guide for initiation to plurilingual approaches: For an equitable and inclusive French language school in Ontario]*. Ottawa: le Centre franco-ontarien de ressources pédagogiques (Le Centre franco). https://downloads.ctfassets.net/cfektv4t16rw/4rqxajQod1gbanWQFZ0RVT/06677b3bb793a03c387e9e9a3a27e8ef/guide-approches-plurilingues-211014.pdf (accessed 27 December 2024)

Masciotra, Domenico & Denise Morel. 2011. *Apprendre par l'expérience active et située: La méthode ASCAR [Learning through Active Situated Experience: The ASCAR Method].* Québec: Presses de l'Université du Québec.

Moore, Danièle. 2019. Conversations autour du plurilinguisme. Théorisation du pluriel et pouvoir des langues [Conversations around plurilingualism: Theorising the plurality and power of languages]. *OLBI Working Papers 10.* 43-64. https://doi.org/10.18192/olbiwp.v10i0.3828

Moore, Danièle & Laurent Gajo. 2009. Introduction. French voices on plurilingualism and pluriculturalism. Theory, significance and perspective. *International Journal of Multilingualism 6*(2). 137–153. https://doi.org/10.1080/14790710902846707

North, Brian. 2007a. The CEFR illustrative descriptor scales. *Modern Language Journal 91*. 655-658. https://doi.org/10.1111/j.1540-4781.2007.00627_3.x

North, Brian. 2007b. The CEFR Common Reference Levels: Validated reference points and local strategies. In Council of Europe (ed.), *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities. Intergovernmental Language Policy Forum*, Strasbourg, 6-8 February 2007, Report, 19-29. Strasbourg: Council of Europe. https://rm.coe.int/090000168069b821 (accessed 27 December 2024)

North, Brian 2014. *The CEFR in practice*. Cambridge: Cambridge University Press.

North, Brian & Piccardo, Enrica. 2016. *Developing illustrative descriptors of aspects of mediation for the Common European Framework of Reference (CEFR)*. Research report. Strasbourg: Council of Europe. https://rm.coe.int/168073ff31 (accessed 27 December 2024)

North, Brian, Enrica Piccardo, Tim Goodier, Daniela Fasoglio, Rosanna Margonis-Pasinetti & Bernd Rüschoff (eds.), 2022. *Enriching 21st century language education: The CEFR Companion Volume in practice*. Strasbourg: Council of Europe. https://rm.coe.int/1680a68ed0 (accessed 27 December 2024)

Pavlovskaya, Irina Y., & Lankina, Olga Y. 2019. How new CEFR mediation descriptors can help to assess the discussion skills of management students: Global and analytical scales. *CEFR Journal - Research and Practice 1*. 33-40. https://doi.org/10.37546/JALTSIG.CEFR1-3

Pearce, Daniel Roy. 2024. *Plurilingual education in a monolingualised nation: Exploring new frontiers in language teaching in Japan*. Cleveland: Multilingual Matters.

Perevertkina, Marina, Korenev, Alexey & Zolotareva, Maria. 2020. Developing classroom mediation awareness and skills in pre-service language teacher education. *CEFR Journal - Research and Practice 2*. 25-39. https://doi.org/10.37546/JALTSIG.CEFR2-2

Piccardo, Enrica. 2005. Complessità e insegnamento delle lingue straniere: ripensare un paradigma. *RILA (Rassegna Italiana di Linguistica Applicata) 37*(2-3). 75-92. https://doi.org/10.1400/57800

Piccardo, Enrica. 2013. Plurilingualism and curriculum design: Towards a synergic vision. *TESOL Quarterly 47*(3). 600-614. https://doi.org/10.1002/tesq.110

Piccardo, Enrica. 2014. *From communicative to action-oriented: A research pathway*. Curriculum Services Canada / Service des programmes d'études Canada https://transformingfsl.ca/en/components/from-communicative-to-action-oriented-a-research-pathway/ (accessed 27.12.2024). French version: *Du communicatif à l'actionnel: un cheminement de recherche*. https://transformingfsl.ca/wp-content/uploads/2014/09/Du-communicatif_a_l_actionnel-Un_cheminement_de_recherche.pdf (accessed 27 December 2024)

Piccardo, Enrica. 2017. Plurilingualism as a catalyzer for creativity in superdiverse societies: A systemic analysis. *Frontiers in Psychology 8*. https://doi.org/10.3389/fpsyg.2017.02169

Piccardo, Enrica. 2018. Plurilingualism: Vision, conceptualization, and practices. In Peter Pericles Trifonas & Themistoklis Aravossitas. (eds.), *Springer international handbooks of education: Handbook of research and practice in heritage language education,* 207-225. New York, NY: Springer. https://doi.org/10.1007/978-3-319-38893-9_47-1

Piccardo, Enrica. 2019. Rethinking plurality in our liquid societies. In Francis Bangou, Monica Waterhouse & Douglas Fleming. (eds.). *Deterritorializing language, teaching, and learning: Deleuzo-Guattarian perspectives on second language education,* 58-86. Leiden/Boston: Brill/Sense https://brill.com/view/title/54763 (accessed 27 December 2024)

Piccardo, Enrica. 2020. The Common European Framework of Reference (CEFR) in language education: Past, present, and future. *TIRF Language Education in Review (LEiR) Series.* https://www.tirfonline.org/wp-content/uploads/2021/10/LEiR_CEFR.pdf (accessed 27 December 2024)

Piccardo, Enrica. 2024. ELT is full of disasters. Teacher talking time: The learn YOUR English podcast. Episode 102: 25.09.2024. https://podcasts.apple.com/ca/podcast/teacher-talking-time-the-learn-your-english-podcast/id1451013687?i=1000670710398 (accessed 27 December 2024)

Piccardo, Enrica, Aline Germain-Rutherford & Geoff Lawrence. (eds.) 2022b. *The Routledge handbook of plurilingual language education*. London and New York: Routledge.

Piccardo, Enrica, & Danielle Hunter. 2017. Settlement, integration and language learning: Possible synergies. A task-based, community-focused program from the Region of Durham (Ontario, Canada). In: Jean-Claude Beacco, Hans-Jürgen Krumm, David Little & Philia Thalgott (eds.), *The linguistic integration of adult migrants,* 175-180. Strasbourg: Council of Europe. https://rm.coe.int/09000016806ae621 (accessed 27 December 2024)

Piccardo, Enrica, & Langé, Gisella. (eds.). 2023. *La classe plurilingue e approccio orientato all'azione*. Milano: Sanoma.

Piccardo, Enrica, Geoff Lawrence, Aline Germain-Rutherford, & Angelica Galante. 2022b. *Activating linguistic and cultural diversity in the language classroom*. New York: Springer.

Piccardo, Enrica, & Brian North. 2019. *The action-oriented approach: A dynamic vision of language education.* Multilingual Matters.

Piribauer, Gerda, Ute Atzlesberger, Irmgard Greinix, Thomas Ladstätter, Franz Mittendorfer, Helmut Renner & Belinda Steinhuber. 2015. *Plurilingualism: Designing and implementing plurilingual oral exams: Framework for the Austrian upper secondary level oral leaving examination at colleges for higher vocational education*. Vienna: CEBS. https://www.cebs.at/wp-content/uploads/2019/05/Austrian_framework_plurilingual_oral_exams-Druckausgabequalit-1.pdf (accessed 27 December 2024)

Prasad, Gail. 2014. Portraits of plurilingualism in a French International School in Toronto: Exploring the role of the visual methods to access students' representations of their linguistically diverse identities. *Canadian Journal of Applied Linguistics 17*(1). 55-71. https://journals.lib.unb.ca/index.php/CJAL/article/view/22126 (accessed 27.12.2024)

Prasad, Gail. 2015. Beyond the mirror towards a plurilingual prism: Exploring the creation of plurilingual 'identity texts' in English and French classrooms in Toronto and Montpellier. *Intercultural Education 26*(6). 497-514. https://doi.org/10.1080/14675986.2015.1109775

Puren, Christian. 2004. De l'approche par les tâches à la perspective co-actionnelle [From task-based language teaching to the co-action-oriented approach]. *Les cahiers de l'APLIUT. Pédagogie et Recherch 23*(1). 10-26. https://journals.openedition.org/cahiersapliut/3416?lang=en (accessed 27.12.2024)

Puren, Christian. 2009. La nouvelle perspective actionnelle et ses implications sur la conception des manuels de langue. [The new action-oriented perspective and its implications for the conceptualisation of language teaching course books]. In Marie-Laure Lions-Olivieri & Philippe Liria (eds.) *L'approche actionnelle dans l'enseignement des langues. Douze articles pour mieux comprendre et faire le point [The action-oriented approach in language teaching: Twelve articles to better understand the state-of-the-art],* 119-137. Paris: Difusión-Maison des langues.

Reimann, Daniel & Andrea Rössler. (eds.) 2013. *Sprachmittlung im Fremdsprachenunterricht [Linguistic mediation in foreign language teaching].* Tübingen: Narr Francke Attempto.

Richer, Jean-Jacques. 2009. Lectures du Cadre: continuité ou rupture? [Readings of the CEFR: Continuity or rupture?]. In Marie-Laure Lions-Olivieri & Philippe Liria (eds.) *L'approche actionnelle dans l'enseignement des langues. Douze articles pour mieux comprendre et faire le point [The action-oriented approach in language teaching: Twelve articles to better understand the state-of-the-art],* 13-48. Paris: Difusión-Maison des langues.

Rosen, Evelyne (ed.). 2009. La perspective actionnelle et l'approche par les tâches en classe de langue [The action-oriented approach and task-based teaching in the language classroom]*. Le Français dans le Monde, Recherches et Applications*, numéro spécial 45.

Sánchez Cuadrado, Adolfo. 2022. *Mediacion en el aprendizaje de lenguas - estrategias y recursos: Investigacion y transferencia ELE [Mediation in the teaching of languages – Strategies and resources*. Research and transfer to Spanish as a foreign language]. Madrid: Anaya. https://www.casadellibro.com/libro-mediacion-en-el-aprendizaje-de-lenguas-estrategias-y-recursos/9788469891605/12830189 (accessed 27.12.2024)

Savski, Kristof. 2019. Putting the plurilingual/pluricultural back into CEFR: Reflecting on policy reform in Thailand and Malaysia. *The Journal of Asia TEFL 16*(2). 644–652. http://dx.doi.org/10.18823/asiatefl.2019.16.2.14.644

Savski, Kristof. 2020. Local problems and a global solution: Examining the recontextualization of CEFR in Thai and Malaysian language policies. *Language Policy, 19,* 527–547. https://doi.org/10.1007/s10993-019-09539-8

Schleiss, Myriam & Margrit Hagenow-Caprez. (2017). fide – On the way to a coherent framework. In Jean-Claude Beacco, Hans-Jürgen Krumm, David Little & Philia Thalgott (eds.), *The Linguistic Integration of Adult Migrants* [*L'intégration linguistique des migrants adultes*] (pp. 169–174). Berlin: De Gruyter.

Soars, John & Liz Soars. 1987. *Headway intermediate*. Oxford: Oxford University Press.

Stathopoulou, Maria. 2015. *Cross-language mediation in foreign language teaching and testing.* Bristol: Multilingual Matters.

Stathopoulou, Maria. 2020. The new CEFR descriptors for the assessment of written mediation: Exploring their applicability in a local context in an effort towards multilingual testing. *CEFR Journal: Research and Practice 2*. 40-78. https://doi.org/10.37546/JALTSIG.CEFR2-3

Stathopoulou, Maria, Phyllisenne Gauci, Magdalini Liontou & Sílvia Melo-Pfeifer. 2023. *Mediation in teaching, learning and assessment (METLA): A teaching guide for language educators.* Graz: Council of Europe. https://www.ecml.at/Portals/1/6MTP/project-stathopoulou/documents/METLA-teaching-guide-EN.pdf (accessed 27 December 2024)

Steinhuber, Belinda. 2022. Implementing plurilingual oral exams and plurilingual lessons in Austrian upper secondary vocational colleges. In Brian North, Enrica Piccardo, Tim Goodier, Daniela Fasoglio, Rosanna Margonis-Pasinetti & Bernd Rüschoff (eds.), *Enriching 21st century language education: The CEFR Companion Volume in practice,* 109-116. Strasbourg: Council of Europe Publishing. https://rm.coe.int/1680a68ed0 (accessed 27 December 2024)

Thornbury, Scott. 2016. Communicative language teaching in theory and practice. In Graham Hall. (ed.), *The Routledge handbook of English language teaching,* 224-237. London/New York: Routledge. https://doi.org/10.4324/9781315676203

Trim, John L. M. 2007. The CEFR in relation to the policy aims of the Council of Europe. In Council of Europe (2007) (ed.) *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities.* Intergovernmental Language Policy Forum, Strasbourg, 6-8 February 2007, Report, 50-51. Strasbourg: Council of Europe. https://rm.coe.int/090000168069b821 (accessed 27 December 2024)

van Lier, Leo. 2000. From input to affordance: Social-interactive learning from an ecological perspective. In James P. Lantolf (ed.) *Sociocultural theory and second language learning,* 245-259. Oxford: Oxford University Press.

van Lier, Leo. 2004. *The ecology and semiotics of language learning*. Dordrecht: Kluwer Academic.

van Lier, Leo. 2007. Action-based teaching, autonomy and identity. *Innovation in Language Teaching and Learning 1*(1). 46-65. https://doi.org/10.2167/illt42.0

van Lier, Leo. 2010. The ecology of language learning: Practice to theory, theory to practice. *Procedia - Social and Behavioral Sciences 3*. 2-6. https://doi.org/10.1016/j.sbspro.2010.07.005

Webb, Noreen M. 2009. The teacher's role in promoting collaborative dialogue in the classroom. *British Journal of Educational Psychology 78*(1). 1-28. https://doi.org/10.1348/000709908X380772

5. Biography

Brian North has been an English teacher, project coordinator, course director, development manager and researcher, spending his career at the interface between research, practical implementation of innovation, and quality management. After implementing a 'can-do' curriculum for different languages in the schools of the Eurocentres Foundation, he was chair of Eaquals, the international quality assurance association for language programmes from 2005 to 2010. He coordinated the 1991 intergovernmental Symposium that recommended the *Common European Framework of Reference for Languages* (CEFR) and *European Language Portfolio* (ELP), for which he developed the levels and descriptors in his PhD (*The Development of a Common Framework Scale of Language Proficiency* 2000: Peter Lang) as part of a Swiss National Science Foundation project. He co-authored the CEFR, the prototype ELP, the Manual for relating examinations to the CEFR, and Eaquals' CEFR Core Inventories for English and French. More recently he coordinated the 2013–20 project that produced the *CEFR Companion Volume* and is main editor of *Enriching 21st century Language Education: The CEFR Companion Volume, Examples from Practice* (2022, Council of Europe). Other publications include: *The CEFR in Practice* (2014: Cambridge); *Language Course Planning* (co-authored: 2018: Oxford), and *The Action-oriented Approach* (with E. Piccardo, 2019: Multilingual Matters).

Appendix

API Mission statement

API Education Forum is an international community of academics, teachers, teacher educators, curriculum developers, administrators, and policy makers in the broad field of language and literacy education. This community is dedicated to the promotion of **Action-oriented, Plurilingual and Intercultural** (API) **Education** at primary, secondary and tertiary levels as well as in adult education, including the integration of migrants within these sectors. API's work is grounded in promoting and protecting linguistic and cultural diversity, which is crucial to equitable, inclusive, quality education. It is also vital for developing mutual understanding and overcoming barriers in order to collaborate and thrive together in today's complex world.

API aims to provide a forum to bring together the latest research in language education, innovative projects across different contexts, and cutting-edge resources in order to support policy development and implementation. A crucial resource for the realization of API Education is the Common European Framework of Reference for languages (CEFR) project. This ongoing and open-ended project includes the CEFR 2001 publication and its 2020 extended edition – the CEFR Companion Volume.

The Forum aims to become an interactive, collaborative space for members to:

- discuss aspects of action-oriented, plurilingual and intercultural (API) education;
- reflect on the challenges and opportunities they present at the classroom, institutional and policy-making levels;
- share practical examples of API implementation and other resources;
- draw attention to and/or collaborate in both relevant academic research and also development projects;
- create, share and disseminate resources, including exemplar scenarios/tasks and/or training modules;
- support the further development of teacher (pre-service and in-service) competences for API education.

Activities within the API Education Forum aim to contribute to the further development of the following key areas within the field of language education:

- awareness-raising and networking concerning projects and initiatives relevant to the CEFR/CV and API education;
- design and development of curricula and resources that promote inclusivity, develop student agency, and foster linguistic and cultural repertoires;
- coverage of the CEFR project and API education in pre-service and in-service professional development;
- development of teaching and learning materials;
- promotion of languages across the curriculum, the development of partial competences in multiple languages, and the use of multiple languages in the classroom;
- support for language education and education for democracy, human rights, and the rule of law, and decolonialization;
- exploration of the potential of multimodality, technology, and large language models (such as ChatGPT) for API education;
- investigation of the role of mediation in plurilingual education, not only as a tool for the language learning classroom, but also as a cross-curricular resource;
- encouragement of communities of practice for teachers working with API education at a local and

regional level;

- collaboration between institutions and educational sectors at the local and regional level.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Understanding the perspective of plurilingual assessment in teaching English at tertiary level in Ukraine

Viktoriia Osidak, Taras Shevchenko National University of Kyiv, Ukraine
Karin Vogt, Heidelberg University of Education, Germany
Maryana Natsiuk, Taras Shevchenko National University of Kyiv, Ukraine

https://doi.org/10.37546/JALTSIG.CEFR7-2

Considering the need for improving assessment instruments that measure language proficiency of plurilingual learners in the foreign language classroom, this article investigates the potential of plurilingual assessment in language education in the Ukrainian context. For this purpose, a developmental project has been carried out engaging several universities. The CEFR and its Companion Volume (CEFR/CV) were used as foundational documents to understand goals of language education and approaches to teaching, learning and assessment. The project was implemented in three phases. During phase 1, a focus was put on the analysis of contributed samples of tests (14 tests comprising 70 assessment tasks) to identify prevailing approaches to language assessment at tertiary level in the Ukrainian context. Most of the contributed assessment tasks (87%) were in English, with a smaller portion (12.8%) both in Ukrainian and English, with 11% out of 12.8% being translation tasks. No assessment tasks were in or more (2+) languages. Phase 2 aimed at empowering the teachers (n=16) with the procedures and assessment instruments to facilitate the implementation of plurilingual assessment in teaching English. Phase 3 collected teacher feedback on proposed changes to language assessment in teaching English using a questionnaire and reflection logs. The outcome of the workshops suggested that plurilingual assessment reflects real-life and professional situations that students can find themselves in but does not seem to represent common practice in the teaching context. In addition, participating teachers indicated that plurilingual assessment is of great relevance to the learning goals of their courses.

Keywords: assessment practices, linguistic repertoire, plurilingualism, plurilingual language assessment, CEFR/CV

1 Introduction

English is increasingly used worldwide as a language of communication and education. In educational contexts, English is often taught as a subject in schools and frequently serves as a medium of instruction in universities. Learners of English are typically emergent multilinguals, for whom English becomes their third language (L3) after their home language(s) (L1) and a second language (L2), which may be acquired through schooling (Sridhar and Sridhar 2018). In the Ukrainian educational context, learners of English are often bilingual in Ukrainian and Russian or another regional minority language. Consequently, English becomes their L3 when their home language and the school language differ.

Thus, recent developments in language teaching and learning when English is their L3 for most learners, make it necessary "to recognise the language ability that language learners already have when learning English" (Seed 2020: 5) and use the knowledge of other languages as a tool in learning English (Seed 2020: 6).

New developments in teaching and learning English in the last few decades have responded to a more diversified linguistic reality in societies (Cummins 2008; Duarte and Gogolin 2013; Tsagari et al. 2023). For teaching English as a foreign language (EFL), the development of multilingualism/plurilingualism and plurilingual assessment is essential, given its importance as an international language. Yet, little research has been done in order to help teachers to implement this multilingual turn in EFL or English as an additional language (EAL) classrooms in Ukraine. Duarte and Günther-van der Meij (2020) mainly attribute this to the fact that a monolingual norm is commonly applied to the understanding of language, learners and the learning process. In addition, the European policy agenda (L1+2 (European) languages) is targeted at promoting additive multilingualism at school level, treating languages as separate entities. As a result, many teachers base their classroom activities on language separation practices. Contrary to classroom practices, learners mobilise their entire linguistic resources in real-life contexts in order to accomplish tasks for personal and communicative purposes (COE 2020).

The current practice of keeping languages apart rather than embracing the full linguistic repertoire of students, presents a dilemma for teachers. Studies e.g., by Duarte and Günher-van der Meij (2020) evidence that language teachers often express positive attitudes towards plurilingualism. Yet, some studies carried out in European and Asian educational contexts indicate that language teachers struggle to implement these attitudes in their instructional practice (e.g., Bisai and Singh 2018; Duarte and Günther-van der Meij 2020). While teachers recognise the value of multilingualism, they may lack clear guidance on how to integrate it effectively into their instructional strategies. This ambivalence highlights the need for greater support without which teachers may feel uncertain about how to assess students' language skills in a way that acknowledges and values their diverse linguistic backgrounds. As a result, students may not have the opportunity to fully demonstrate their entire linguistic repertoire, and the potential benefits of plurilingualism in the classroom may remain untapped. For foreign language teaching and assessment, this means considering multilingual resources already present in diverse learning groups.

2 Literature review

2.1 Terminology

The Common European Framework of Reference (CEFR) (COE 2001) and its updated policy document, the CEFR/CV (COE 2020), make a distinction between *multilingualism* and *plurilingualism*. Multilingualism is defined as the coexistence of different languages at the social or individual levels while plurilingualism as the dynamic and developing linguistic repertoire of an individual user/learner (COE 2020: 28). A person is seen as a social agent, using their language repertoire in order to accomplish a task or an action (Piccardo and North 2019). In educational settings, plurilingualism takes an individual perspective that aims to capture the holistic and dynamic nature of the individual learner's linguistic repertoire as it develops through life (COE 2001: 168). A plurilingual learner has a "single, interrelated, repertoire that they combine with their general competences and various strategies in order to accomplish tasks" (COE 2020: 30). In this context, following the CEFR descriptive framework and the action-oriented approach, the focal point of the learning and teaching process is the collaborative creation of meaning through interaction (COE 2020). From this standpoint, plurilingual language assessment takes a perspective that recognises the interconnectedness of languages in an individual's repertoire and considers the holistic and dynamic nature of language use across multiple languages. In essence, it aims to assess overall communicative competence, considering how languages are integrated and used together.

2.2 A multilingual turn in assessment?

Although assessment is an inherent part of the education process and multilingual education has been discussed for several decades, little attention has been paid to multilingualism or plurilingualism in assessment and the much-cited multilingual turn (Conteh and Meier 2014) has not become a reality in

language assessment yet. As a result, plurilingual learners are still being predominantly assessed in each language separately (Chalhoub-Deville 2019; De Backer et al. 2020; Tsagari et al. 2023). According to Choi et al. (2022), the current prevailing monolingual approach to language assessment that conceptualises languages as separate entities fails to acknowledge complex communicative practices of plurilinguals and their ability to draw on their diverse linguistic repertoire and are invalid in terms of assessing what plurilingual learners know or can do (Choi et al. 2022: 333). Furthermore, in Bisai and Singh's view (2018: 309), assessment from a monolingual standpoint fails to capture the reality of the EFL classroom. There is a shared understanding that language assessment tasks should provide learners with opportunities to demonstrate their relevant language skills by observing performance on relevant and authentic tasks. Gorter and Cenoz (2017) advocate that if teaching is to consider plurilingual concerns, assessment practices should follow suit.

The integration of plurilingual assessment has always been a challenge in many respects: operationalising a construct for authentic assessment tasks, and providing reliable scoring are among plurilingual assessment concerns. One of the reasons for such a challenge is that plurilingual assessment tasks should be personalised as they "would depend on the contexts that each plurilingual, pluricontextual language learner finds themselves in" (Seed 2020: 9). The same idea is reiterated by Bisai and Singh (2018) who argue that the language resources mobilised by plurilinguals are individualised, dynamic, and contextualised. To meet the requirements of plurilingual assessment, assessment should be multimodal, integrated, fluid, and ongoing, and these qualities are largely compatible with alternative and formative assessment (Gorter and Cenoz 2017; Poehner and Inbar-Lourie 2020; Seed 2020).

2.3 Plurilingual assessment of English as a Foreign Language

In recent years, the question of how plurilingual assessment can be organised has received increasing attention. Seed (2020) specifies the framework of assessment in plurilingual situations into four broad constructs that can capture individuals' plurilingual abilities in four different ways. In essence, the framework distinguishes between assessment of language proficiency in one or several named language(s), assessment of content knowledge and the assessment of plurilingual competence that includes learners' competence of both languages known and only partially known.

The focus of this paper is on plurilingual assessment in foreign language education, which relates to assessment in one named language such as English with both input and output in that language. Seed (2020) argues that language tests, even if they are monolingual, should be considered as integral components of a broader multilingual language profile that a person can demonstrate in multilingual situations (Seed 2020: 10; Seed and Holland 2020). Schissel et al. (2018) found that tasks that integrate multilingual reading materials result in better performance by plurilingual participants compared to English-only tasks. Therefore, instances of other languages during assessment should be taken as evidence of assistance in accomplishing a task (communication) successfully. The findings, suggesting that incorporating multilingual resources in language assessment design can enable language learners to exhibit more advanced writing skills and higher-order thinking abilities, may become a valuable pedagogical implication for plurilingual assessment in the EFL classroom.

Flexible plurilingual assessment methods that recognise learners' (partial) proficiency in multiple languages have recently received much attention. Such assessment is based on the idea that learners are disadvantaged if they are not allowed to build on their whole linguistic repertoire (De Backer et al. 2020). In fact, plurilingual assessment acknowledges the different skills that plurilinguals require, such as the use of other languages and the role of their cross-linguistic and metalinguistic skills to complete a test task (Lopez et al. 2017).

According to North and Piccardo (2016, 2017) and Stathopoulou (2020), people communicate using a combination of different languages, making it important for language users to develop the ability to mediate cross-linguistically. Mediation as a common cross-linguistic activity involves moving between

different languages with the purpose to explain, clarify, interpret, summarise, or convey the main points of a text to someone else (North and Piccardo 2016, 2017). Mediation always occurs in a social context (public, academic, and professional) and is a purposeful activity that language users engage in when there is a communication gap (COE 2020). Therefore, a test that combines two or more languages can be a solution for assessing English in a multilingual context. In this regard, the CEFR/CV provides scales for different aspects of mediating a text (including literature), mediating concepts, and mediating communication (COE 2020: 91-122). In addition, the CEFR/CV provides scales for signposting different aspect of a plurilingual repertoire in a task: Scales for *Building on plurilingual repertoire* and *Building on pluricultural competence; Plurilingual comprehension* (COE 2020: 124-128). North and Piccardo (2023) highlight that descriptors are important tools that can support teachers and learners in several respects. The descriptors can empower teachers in their desire to promote a plurilingual approach to teaching and assessment; suggest real world-oriented classroom tasks and become an indicator of students' performance etc. Likewise, descriptors can also help learners become aware of their plurilingual repertoire, and demonstrate the purpose of the activity.

Despite the availability of CEFR/CV scales for mediating texts and concepts and building on plurilingual competence, there remains a gap in the practical implementation of plurilingual assessment. Specifically, current assessments of English often do not create opportunities for learners of English to engage with their whole linguistic repertoire in plurilingual contexts effectively. Thus, our research aims to address this gap by developing a test that incorporates multiple languages, and leveraging CEFR/CV descriptors to support a plurilingual approach to language assessment. To effectively address this goal, the paper will investigate Ukrainian Higher Education Institution (HEI) language teachers' assessment practices and strategies regarding plurilingual assessment. As the project involved a follow-up workshop, its further objective was to empower university teachers with knowledge about plurilingualism in language education and assessment strategies designed to facilitate the implementation of plurilingual assessment in teaching English to pre-service teachers and students majoring in Linguistics. Therefore, the following research questions have been formulated:

1. To what extent are the samples of assessments from Ukrainian universities plurilingual?
2. What strategies were employed to develop plurilingual tasks to assess students' proficiency in English?
3. What strategies were employed to tailor descriptors selected from the CEFR/CV relevant to the local context?
4. How do HEI language teachers based in Ukraine evaluate the proposed changes to existing language assessment?

3 Research Methodology

3.1 Participants

The data was obtained from two sets of participants. Convenience sampling was used for the purpose of this developmental project (Dörnyei 2007). Although we were aware of the disadvantages of convenience sampling such as a possibly imbalanced sample, convenience sampling was used due to ease and the participants' voluntary agreement to commit their time and effort to the research goals, which was especially crucial due to the war-related circumstances in Ukraine. The first group, 16 University English teachers from National University *Yuri Kondratyuk Poltava Polytechnic*, volunteered to participate in the workshop training and complete the online questionnaire. In addition, five of these 16 teachers volunteered to fill in the reflection logs. All participants gave written informed consent to their participation in the study, and all data collected were anonymised.

3.2 Method

A mixed-methods approach was used to collect data to answer the research questions of the study. Using a mixed-method study design has a number of advantages over a single method in educational research, especially when exploring a new phenomenon (Cohen et al. 2007; Dörnyei 2007). By applying different methods of collecting data, including analysis of the assessment tasks, a small-scale questionnaire survey and reflection logs, we were able to achieve a more comprehensive understanding of the matter from multiple perspectives. The chosen approach aimed at triangulating data from these different sources, enabling us to answer our research questions while also supporting evidence for drawing conclusions. (Cohen et al. 2007). In this light, quantitative methods (a small-scale questionnaire survey and descriptive statistics of the data) were used to collect explicit numerical evidence (Creswell 2009) about existing assessment practices and strategies employed in developing plurilingual assessment tasks. Descriptive statistics (the mean) was used to establish types of assessment tasks by calculating the percentage and to identify a set of strategies related to developing plurilingual assessment tasks and customising the descriptors to the local context. In addition, descriptive statistics (percentage) was calculated to interpret the data collected by the questionnaire. Among the strategies of inquiry of a qualitative method, a reflection log was employed to arrive at a 'thick description' (Younas et al. 2023) of the participants' experience and the development in their assessment practice.

3.3 Project design

The project framework includes three subsequent phases: Understanding of the local context, awareness and engagement, and evaluation (see Table 1).

Table 1. ***Phases of the project design***

Project design		
Phase #	Description of the phase	Activities
Phase 1: Understanding of the local context RQ 1: To what extent are the samples of assessments from Ukrainian universities plurilingual?"	Collecting and analysing assessments from Ukrainian universities: 14 sample tests consisting of 70 tasks.	Collaborating with colleagues from different HEI; Reflective practice
Phase 2: Awareness and engagement RQ 2: What strategies were employed to develop plurilingual tasks to assess students' proficiency in English? RQ 3: What strategies were used to customise the descriptors to the local context?	Workshop 1 (90 min): (16 participants) Input relating to the basic CEFR/CV related concepts: multilingualism vs plurilingualism, language competence, partial competence, native-speaker standard, language portraits and individual language profiles, linguistic repertoires, monolingual/ multilingual approaches to language teaching and assessment, cross-linguistic mediation etc.	Participating in training
	Workshop 2 (90 min): (16 participants) Input relating to plurilingual assessment strategies: • Discussing plurilingual assessment strategies • Adapting assessment tasks to plurilingual contexts • Presentation of adapted assessment tasks • Discussing descriptors • Selecting and customising descriptors • Presentation of adapted descriptors	Brainstorming; Group discussion; Collaborating in breakout rooms

Project design		
Phase #	Description of the phase	Activities
Phase 3: Evaluation RQ 4: How do HEI language teachers based in Ukraine evaluate the proposed changes to existing language assessment?	Mixed methods Collecting teacher feedback: • Reflection logs (5 participants) • Online survey (16 participants)	Reflective practices

In Phase 1, colleagues from three universities contributed tests used at their departments to assess students' proficiency in English. The analysis of the assessment tasks was carried out with the purpose to understand to what extent the samples of assessments were plurilingual. To this end, the collected assessment tasks were scrutinised against the following aspects 1) the targeted competences, 2) whether a test enables students to demonstrate their plurilingual comprehension and/ or build on their plurilingual repertoire; 2) target language(s) of input and output; 3) assessment types.

The awareness and engagement phases included two online workshops using Zoom. The workshops lasted 90 minutes each and were held within one week. The purpose of workshop 1 (Awareness) was to familiarise the participants with the key concepts related to the field of multilingualism/ plurilingualism (see Table 1) in order to establish a common knowledge base. It also helped to understand fundamental concerns in multilingual/plurilingual language education to eliminate possible misinterpretations. In addition, workshop 1 was designed to give all the participants the possibility to analyse their local contexts and consider whether plurilingual assessment tasks are compatible with their existing assessment framework.

Workshop 2 (Engagement) was aimed at engaging the teacher participants to demonstrate their competency in modifying assessment tasks to the plurilingual context, selecting the descriptors from the CEFR/CV and customising them to the modified tasks. For this purpose, the workshop included several steps.

First, the teachers were invited to analyse the original assessment tasks. They collectively offered suggestions as to how a monolingual task can be adapted to a plurilingual context (see Table 2).

Table 2. ***An example of a task modification during the workshop (modifications added in blue).***

Original task: Plan a group vacation The sources are given in English.	**Modified task: Plan a group vacation** The sources are given in English and Ukrainian
As a group, decide on a budget for your vacation and select a destination that everyone is interested in. Analyse travel brochures, online websites, and other sources of information to find the best options for your group vacation. Look for destinations that offer activities and attractions that match the interests and preferences of everyone in the group. Choose two or three destinations that you think would be the most suitable for your group vacation, and present your analysis to the class or in a video.	As a group, decide on a budget for your vacation and select destinations that everyone is interested in. Analyse travel brochures, online websites, and other sources of information in two languages that popularise different destinations in Britain and in Ukraine to find the best options for your group vacation. Look for destinations that offer activities and attractions that match the interests and preferences of everyone in the group. Choose two destinations (one in Britain and one in Ukraine) that you think would be the most suitable for your group vacation, and present your analysis to the class or in a video in English.

Next, teachers were invited to collaborate in breakout rooms, forming groups of four. Their collective objective was to propose plurilingual strategies aimed at adapting assessment tasks collected during

Phase 1. The teachers engaged in collaborative discussions that contributed to co-constructing knowledge on designing plurilingual assessment tasks. Subsequently, each group in turn showcased the outcomes by presenting the modified task. Finally, the teachers submitted the modified assessment tasks to the authors for further analysis.

The next step of workshop 2 included discussing and localising the descriptors relevant to the task using the CEFR/CV as a benchmark. The teachers worked following the same pattern: discussing descriptors – collaboration in breakout rooms – presenting descriptors – submitting the outcome of collaborative product to the authors for further analysis. The added descriptors to the tasks drew on the following scales: Building on plurilingual comprehension, pluricultural competence and mediation (see Table 3). After compiling a list of descriptors from the CEFR/CV, the possibilities of adjusting those descriptors were discussed.

Table 3. ***Relevant descriptors from the CEFR/CV***, descriptors for the original task are in black; strategies are in blue; added descriptors to a modified task are in green.

Reading for orientation	B1+	Can scan through straightforward, factual texts in magazines, brochures or on the web, identify what they are about and decide whether they contain information that might be of practical use (COE 2020: 56).
Sustained monologue: Putting a case	B1	Can briefly give reasons and explanations for opinions, plans and actions (COE 2020: 64)
Oral production: Addressing audience	B1	Can give a prepared straightforward presentation on a familiar topic within their field which is clear enough to be followed without difficulty most of the time, and in which the main points are explained with reasonable precision (COE 2020: 66).
Overall mediation	B1	Can convey information given in clear, well-structured informational texts on subjects that are familiar or of personal or current interest, although lexical limitations cause difficulty with formulation at times (COE 2020: 92).
Planning	B1	Can work out how to communicate the main point(s) they want to get across, exploiting any resources available and limiting the message to what they can recall or find the means to express (COE 2020: 69)
Collaborating in a group	B1+	Can collaborate on a shared task, e.g., formulating and responding to suggestions, asking whether people agree, and proposing alternative approaches (COE 2020: 111)
Processing texts in speech	B1	Can summarise simply (in Language B, namely English- our addition) the main information content of straightforward texts (in Language A, namely Ukrainian – our addition) on familiar subjects (e.g., a short record of an interview, magazine article, travel brochure) (COE 2020: 101).
Building on pluricultural repertoire	B1	Can explain features of their own culture to members of another culture or explain features of the other culture to members of their own culture (COE 2020: 125)
Plurilingual comprehension	B1	Can deduce the message of a text by exploiting what they have understood from texts on the same theme in different languages (e.g., news in brief, museum brochures, online reviews) (COE 2020: 126).

After discussing the suggested descriptors, the participants were given the task to locate the descriptors for the plurilingual tasks modified in Phase 1 and then to customise the selected descriptors. To complete this task, the participants worked in groups of four in breakout rooms. The CEFR/CV (COE 2020) served as reference. The presented results of a collaboration demonstrate that allotted time in breakout rooms was not enough to locate the descriptors and to offer modifications to them. Therefore, it was decided that the groups required more time to finalise the descriptors. Thus, the groups were offered to submit their final descriptors to the authors within 5 days.

Phase 3 collected teachers' views on proposed changes to language assessment in teaching English. For this, a structured online questionnaire was administered to the participants, and the focus group was asked to fill in the reflection log. The questionnaire was open for three weeks during which the 16 participants of the workshop could submit their responses. The focus groups were asked to submit their answers in a weeks' time.

3.4 Data collection and data analysis

To identify to what extent language assessments in teaching English are plurilingual, we approached universities specialised in preparing pre-service EFL teachers and students majoring in Linguistics. Three universities located in different regions in Ukraine volunteered to contribute tests that are developed by their English teachers and are used by the universities to assess their students' proficiency in English. Altogether, the universities contributed 14 sample tests: 8 tests from University 1; 5 tests from University 2 and 1 test from University 3. This imbalance could lead to overrepresentation or underrepresentation of certain variables across universities by thus potentially distorting findings and limiting the generalisability of the conclusions. Consequently, the skewed sample necessitates caution in interpreting the results. Despite this limitation, it was expected that the collected assessment tasks could provide us with insights into the most typical assessment activities used for evaluating the language proficiency of pre-service EFL teachers' and students majoring in Linguistics.

Then, the collected tests were analysed using descriptive statistics (establishing frequencies) in order to define 1) the targeted skills, 2) whether a test enables the students to demonstrate their plurilingual comprehension and/ or build on their plurilingual repertoire; 3) language(s) of input and output; 4) assessment types. The summary of the analysis is presented in Appendices A and B.

A structured questionnaire and a reflection log (see Appendix D) were used to collect teachers' views on proposed changes to the existing language assessments. The questionnaire and the reflection logs consisted of questions aligned with the objectives of the workshops (see Table 1) and targeted three main areas 1) the teachers' understanding of the key concepts of plurilingualism in language education; 2) pedagogical practices used in the language classroom and 3) approaches to language assessment. Altogether, the questionnaire comprised 18 items. A five-step Likert scale, ranging from *'1-totally disagree', '2-disagree', '3-undecided' '4-agree', to '5-totally agree',* was employed. The questionnaire was administered online, using Google Forms among 16 participants immediately after the two workshops. To encourage participants to express their genuine perceptions of the workshop content, all answers were kept anonymous. Then, the frequency for each response was recorded and data were presented in percentages.

The reflection log (11 items) was used to arrive at an in-depth picture of the participants' perceptions of the workshops. Reflective practices in educational context promote teacher critical thinking, and raise awareness about their surrounding and context (Hashim and Yusoff 2021) The data analysis was guided by the exploratory nature of the study and content analysis to ensure valid inferences from the content of textual data (De Wever et al. 2006). Pre-ordinate categorisation was used (Cohen et al. 2007), which means that the authors identified three main categories devised from the areas of their interest in advance. Consequently, the teachers' reflections were analysed according to these categories of keywords: (1) the participants' understanding of plurilingualism in language education, (2) language classroom practices, and (3) the approaches to language assessment. In this light, the codes in this

part of the study were deductive. From the start, coding involved meticulous reading and annotating each teacher reflection material. Then, significant responses were tied to the relatable categories and analysed using an interpretive perspective (Cohen et al. 2007). Finally, the integration and merging of the statistical analysis of the quantitative data and interpretation of qualitative data took place.

4 Results

Research question 1 explored to what extent the samples of assessments from Ukrainian universities were plurilingual. The collected tests measure English proficiency of pre-service teachers and students majoring in Linguistics, targeting English for communication and professional purposes (communication and grammar, 1st to 4th years of study, Bachelor programme) at CEFR levels B1+ to C1 and tasks that target academic English/ English for professional purposes (Master programme) at CEFR levels C1-C1+. The test analysis demonstrates that all 14 tests are characterised by a summative test design. The test from University 3 has a built-in progression through the course. The tests are mainly monolingual, in English. All 14 tests target at an ideal native-speaker language use. In addition, 13 tests include two assessment parts – written and oral and consist of four to six assessment tasks. Altogether, 14 tests include 70 assessment tasks. Language competence is assessed by measuring proficiency in several skills: reading, writing, mediation, speaking, interaction and language functions: grammar and vocabulary. These mostly discrete-point tests do not include tasks which assess listening skills.

Most of the assessment tasks (87%) are in one named language – English. Nine assessment tasks (12.8%) are in two languages, namely Ukrainian and English. Eight (11%) of these tasks are translation tasks: three tasks (University 1) focus on translating isolated sentences comprising target vocabulary from Ukrainian into English and 5 tasks (University 2) focus on translating a written text from English into Ukrainian. In addition, University 2 includes one task that assesses cross-linguistic mediation by relaying specific information in writing, namely summarising and explaining in English the purpose of a dissertation conducted in Ukrainian. No assessment tasks are in 2+ languages. Table 4 illustrates languages involved in tests to assess language proficiency in English.

Table 4. ***Languages in tests to assess language proficiency in English***

Languages involved in 70 tasks	**n**	**%**
Tasks in one language	61	87%
Tasks in two languages	9	12.8%
Tasks in 2+ languages	0	0%
Tasks in mediation	24	34%
Tasks in mediation in one language, English	15	21%
Tasks in mediation in two languages, English and Ukrainian	9	1.5%
Tasks in translation	8	11%

Research question 2 looked into the strategies that the teachers used to develop plurilingual tasks to assess students' proficiency in English. The participants worked in groups of three or four. Each group modified one or two of the assessment tasks collected in Phase 1. The analysis of the modified tasks demonstrated that the teachers successfully employed several strategies to design assessment tasks that engage students' plurilingual competence (see Tables 5-9). Among such strategies were:

- Communicating written or oral information from Ukrainian to English in writing or speaking.
- Summarising information read or heard in Ukrainian (and English) and its further presentation in speaking or writing in English where changes of discourses or genre of the original text(s) are possible.

- Collating information from different sources in Ukrainian and English in order to produce a written text in English.
- Comparing grammar in students' L1 and English.
- Reflecting on an issue raised in Ukrainian and English cultures.

The overarching objective of these language assessment activities is to foster language contact and raise awareness of linguistic and cultural diversity, particularly for languages like English and Ukrainian. By incorporating elements of different languages, students are encouraged to employ their linguistic repertoire in their L1 and English but also demonstrate a deeper understanding of language dynamics and intercultural communication.

Table 5. ***Modifications of the tasks assessing mediation of a text*** (modifications added in blue) as proposed by group 1.

Original task A	**Modification 1**	**Modification 2**
Read the text *Bilinguals* and write a summary paragraph (10-12 sentences) commenting on the issue raised in the text. To what extent do you share the author's opinion?	Read the two texts *Bilinguals* and *Двомовні з дитинства*[1] and write a summary paragraph in English (20-25 sentences) commenting on the issue raised in the texts. Compare and contrast the ideas discussed in the two texts.	Read the text *Двомовні з дитинства* and write a summary paragraph in English (10-12 sentences) commenting on the issue raised in the texts. To what extent do you share the author's opinion?

Original task B	**Modification**
Read a short text and analyse its communicative message. Identify the main problem that the text introduces and provide a detailed explanation, supported by relevant arguments and examples. Additionally, provide recommendations or potential solutions to the problem discussed in the text.	Read a short text in Ukrainian and analyse its communicative message in English. Identify the main problem that the text introduces and provide a detailed explanation, supported by relevant arguments and examples. Additionally, provide recommendations or potential solutions to the problem discussed in the text.

As can be seen from the examples in Table 5, modifications of the tasks often involved cross-linguistic mediation that included introducing an additional text in Ukrainian, or substituting the text in English with a text in Ukrainian (task b).

1. Bilingual from Childhood (our translation)

Table 6. ***Modifications of the tasks assessing grammar*** (modifications added in blue) as proposed by group 2.

Original task A	**Modification**
Rewrite the sentences by using the compound adjectives. Example: A train which moves fast. – A fast-moving train.	Rewrite the sentences using compound adjectives. Example: A train which moves fast. – A fast-moving train. Then, provide the equivalent sentence in Ukrainian and comment in English on the differences in parts of speech used in the two languages. Consider the different structures and word order in Ukrainian and English.
Original task B	**Modification**
Provide a complete syntactic analysis of the sentence 'People who speak more than one language are fascinating.'	Provide a complete syntactic analysis of the sentence 'People who speak more than one language are fascinating' and compare it with the syntactic structure in Ukrainian. Identify and explain any differences between the syntactic structures of the two languages, taking into consideration the word order and sentence structure.

Table 6 demonstrates that the changes to grammar tasks (paraphrase, syntactic analysis of the sentence) included raising language awareness about the differences in syntactic structures used and included analysis and comparison of linguistic structures in English and Ukrainian. Similar modifications to grammar tasks were offered to task b.

According to the CEFR and the CEFR/CV (COE 2001, 2020), plurilingualism entails communication not only across languages, but also across cultures and contexts. Therefore, the teachers of group 3 modified a speaking on the topic monolingual task by including reflection on and the analysis of problems raised in the task from a cultural perspective (see Table 7).

Table 7. ***Modifications of the tasks assessing speaking*** (modifications added in blue) as proposed by group 3.

Original task	**Modification**
Look at the pictures and explain the problems they illustrate.	Look at the pictures and explain the problems they illustrate. Are these problems common for Ukraine too? Compare and contrast the issue and its solutions in the two contexts.

Group 3 deployed the same pluricultural strategy with regard to the task for assessing translation and the analysis of a creative text. First, the teachers omitted the translation task overall. Instead, modification was offered to the analysis of a creative task, which included analysis of a literary text from a cultural perspective (see Table 8).

Table 8. ***Modifications of the tasks assessing translating a written text and relaying information*** (modifications added in blue) as proposed by group 3.

Original task	Modification
1. Read and translate the extract from 'Philomel Cottage' by Agatha Christie (1, 501 words) (Christie, n. d.). 2. Explain the message presented in the extract from 'Philomel Cottage' (67-68) by Agatha Christie.	Read an extract from 'Philomel Cottage' by A. Christie (1, 501 words) (Christie, n.d.) and present a summarised version of the information contained in the text. Pay close attention to any cultural phenomena or references in the text that may not be properly understood by Ukrainian readers. Additionally, provide explanations or context for any cultural references or vocabulary (comment on at least 3 instances) that may be unfamiliar to Ukrainian readers, and use your knowledge of both cultures to bridge any potential gaps in understanding.

Group 4 proposed adaptation of the monolingual collaborative task by introducing a requirement to work with diverse linguistic contexts (see Table 9). Thus, the modified task engages students with the broader scope of the project. This allows students to get a richer and more comprehensive understanding of the topic by exploring authentic, multilingual resources beyond topic-related materials, enhancing their exposure to real-world language use. It also allows them to draw connections between their L1 and the language they are learning, promoting deeper linguistic and cultural understanding.

Table 9. ***Modification of a task assessing a collaborative group project*** (modifications added in blue) as proposed by group 4.

Original task A	Modification
Prepare a collaborative group project that incorporates the topics, vocabulary, and grammar structures learned throughout the course. The project can take the form of a video, performance, or presentation.	Prepare a collaborative group project that incorporates the topics, vocabulary, and grammar structures learned throughout the course. In addition to the course material, utilise podcasts, interviews, videos, and blogs in other languages that you know (including L1) related to the course topics. The project should be presented in English and can take the form of a video, performance, or presentation. Present a reference list of the sources used.

Research question 3 analysed the strategies used by the teachers to customise selected descriptors. After the analysis of the submitted descriptors, the participants drew on the descriptors for cross-linguistic mediation, descriptors on plurilingual comprehension and building on plurilingual repertoire. In order to adjust these descriptors to their contexts, the teachers used three main strategies, namely removing irrelevant information, adding specific details related to the language of input and output or combining several descriptors. Further, we will exemplify teachers' decisions regarding the choice of the descriptors from the CEFR/CV and comment on strategies employed to customise the descriptors.

Table 10 illustrates selecting and adapting relevant descriptors from the CEFR/CV to the assessment task in mediation.

Table 10 Selecting and adapting relevant descriptors: Read the two texts Bilinguals and Двомовні з дитинства and write a summary paragraph (20-25 sentences) commenting on the issue raised in the texts. Compare and contrast the ideas discussed in the two texts.

Table 10. ***Selecting and adapting relevant descriptors***

Criteria	Level	Descriptors from the CEFR/CV	Adaptation of the descriptors
Relaying specific information in writing	B2+	Can relay in writing (in Language B) the relevant point(s) contained in propositionally complex but well-structured texts (in Language A) within their fields of professional, academic and personal interest (COE 2020: 94).	Can relay in writing (in English) the relevant point(s) contained in both of the propositionally complex but well-structured texts presented in Ukrainian and English.
Processing text in writing	B2+	Can compare, contrast and synthesise in writing (in Language B) the information and viewpoints contained in academic and professional publications (in Language A) in their fields of special interest (COE 2020: 99).	Can compare, contrast and synthesise in writing (in English) the information and viewpoints contained in both of the professional publications (in Ukrainian and English).
Plurilingual comprehension	B2	Can use their knowledge of contrasting genre conventions and textual patterns in languages in their plurilingual repertoire in order to support comprehension (COE 2020: 126).	Can use knowledge of contrasting genre conventions and textual patterns in Ukrainian and English in order to support comprehension
Building on plurilingual repertoire	B2	Can alternate between languages in their plurilingual repertoire in order to communicate specialised information and issues on a subject in their field of interest to different interlocutors (COE 2020: 128).	Can alternate between Ukrainian and English in order to communicate specialised information and issues on a subject

Table 11 illustrates selecting and adapting relevant descriptors from the CEFR/CV to the assessment task in grammar. For assessing students' ability to explain the difference between the syntactic structures in the two languages, the teachers located relevant descriptors in plurilingual comprehension and the explaining data scales. As this scale "refers to the transformation into a verbal text of information found in figures" (COE 2020: 96), the syntactic composition of the sentence may be regarded as graphic data, the choice of the descriptor is seen as justifiable.

Table 11 Selecting and adapting relevant descriptors to the assessment task: Provide a complete syntactic analysis of the sentence 'People who speak more than one language are fascinating' and compare it with the syntactic structure in your L1. Identify and explain any differences between the syntactic structures of the two languages, taking into consideration the word order and sentence structure.

Table 11. ***Selecting and adapting relevant descriptors to the assessment task***

Criteria	Level	Descriptors from the CEFR/CV	Adaptation of the descriptors
Plurilingual comprehension	B2	Can use their knowledge of contrasting genre conventions and textual patterns in languages in their plurilingual repertoire in order to support comprehension (COE 2020: 126).	Can use their knowledge of contrasting syntactic structures in languages (English and students' L1) in their plurilingual repertoire in order to support comprehension.
Explaining data	B2	Can interpret and describe reliably (in Language B) detailed information contained in complex diagrams, charts and other visually organised information (with text in Language A) on topics in their fields of interest (COE 2020: 97).	Can interpret and describe reliably in English detailed information contained in syntactic sentence analysis diagram on syntactic differences in English and a student's L1.

As Table 12 shows, the assessment task with the focus on reflection upon and analysis of problems from a cultural perspective was evaluated using descriptors from mediation scales and building on pluricultural repertoire. To adjust the descriptors from the CEFR/CV to the assessment task, information that specified languages involved in assessment was added, irrelevant information was removed. Considerable adjustments underwent the descriptor in the explaining data in speech or sign scales by removing the information about the type of data and the topic.

Table 12 Selecting and adapting relevant descriptors to the assessment task: Look at the pictures and explain the problems they illustrate. Are these problems common for Ukraine, too? Compare and contrast the issue and its solutions in two countries.

Table 12. ***Selecting and adapting relevant descriptors to the assessment task***

Criteria	Level	Descriptors from the CEFR/CV	Adaptation of the descriptors
Overall mediation	B2+	Can establish a supportive environment for sharing ideas and facilitate discussion of delicate issues, showing appreciation of different perspectives, encouraging people to explore issues and adjusting sensitively the way they express things (COE 2020: 92).	Can share ideas of delicate issues, showing appreciation of different perspectives, adjusting sensitively the way they express things.
Explaining data in speech or sign	B2	Can interpret and describe reliably (in Language B) detailed information contained in complex diagrams, charts and other visually organised information (with text in Language A) on topics in their fields of interest (COE 2020: 97)	Can interpret and describe reliably in English detailed information contained in images.

Criteria	Level	Descriptors from the CEFR/CV	Adaptation of the descriptors
Building on pluricultural repertoire	B2	Can explain their interpretation of the cultural assumptions, preconceptions, stereotypes and prejudices of their own community and of other communities that they are familiar with (COE 2020: 125).	No adjustments
Building on pluricultural repertoire	B2	Can generally interpret cultural cues appropriately in the culture concerned (COE 2020: 125)	No adjustments

Similar strategies were applied to the task assessing relaying information (see Table 13). Two descriptors were left without changes. In addition, two descriptors related to building on pluricultural repertoire scales were combined into one.

Table 13 Selecting and adapting relevant descriptors to the assessment task: Read an extract from 'Philomel Cottage' by Agatha Christie (n.d.) and present a summarised version of the information contained in the text. Pay close attention to any cultural phenomena or references in the text that may not be properly understood by Ukrainian readers. Additionally, provide explanations or context for any cultural references or vocabulary (comment on at least 3 instances) that may be unfamiliar to Ukrainian readers, and use your knowledge of both cultures to bridge any potential gaps in understanding.

Table 13. ***Selecting and adapting relevant descriptors to the assessment task***

Criteria	Level	Descriptors from the CEFR/CV	Adaptation of the descriptors
Overall mediation	B2+	Can convey the main content of well-structured but long and propositionally complex texts on subjects within their fields of professional, academic and personal interest, clarifying the opinions and purposes of speakers/signers (COE 2020: 92).	No adjustments
Expressing a personal response to creative texts (including literature)	B2	Can give a personal interpretation of the development of a plot, the characters and themes in a story, novel, film or play (COE 2020: 106).	Can give a personal interpretation of the development of a plot, the characters and themes in a story.
Facilitating pluricultural space	B2+	Can exploit knowledge of sociocultural conventions in order to establish a consensus on how to proceed in a particular situation that is unfamiliar to everyone involved (COE 2020: 115).	No adjustments

Criteria	Level	Descriptors from the CEFR/CV	Adaptation of the descriptors
Strategies to explain a new concept: Adapting language	B2	Can make accessible for others the main contents of a text on a subject of interest (e.g., an essay, a forum discussion, a presentation) by paraphrasing in simpler language (COE 2020: 119).	Can make accessible for others the main contents of a story by paraphrasing in simpler language and breaking into a series of smaller steps.
Strategies to explain a new concept: Breaking down complicated information		Can make a complicated process easier to understand by breaking it down into a series of smaller steps (COE 2020: 119).	
Building on pluricultural repertoire	B2	Can generally interpret cultural cues appropriately in the culture concerned (COE 2020: 125).	Can interpret cultural cues appropriately in the culture concerned by explaining particular ways of communicating in their own and other cultures
Building on pluricultural repertoire	B2	Can reflect on and explain particular ways of communicating in their own and other cultures, and the risks of misunderstanding they generate (COE 2020: 125).	

Similar to other adapted descriptors, the descriptors to a collaborative group project specify the languages used, provide additional clarifications on cultural aspects and strategies used to complete a project (see Appendix C).

To answer Research question 4, a questionnaire survey and a reflection log were used to investigate the teachers' views on the changes to language assessment in teaching English. The collected data will be presented along the three focal pre-ordinate categories: Understanding plurilingual/multilingual goals in language education, plurilingual pedagogical practices in teaching English, approaches to teaching English. Table 14 presents the results of the survey, gauging teacher understanding of plurilingual/ multilingual goals in language education.

Table 14. ***The results of teacher understanding of plurilingual/ multilingual goals in language education***

Understanding key concepts	n=16
In the English classroom, students should NOT learn about the language as a subject.	Disagree - 25% (4) Undecided - 25% (4) Agree – 43.8 (7) Strongly agree - 6.7% (1)
In the English classroom, students should use the language to co-construct meaning and create a product.	Agree - 75% (12) Strongly agree - 25% (4)
I understand the difference between plurilingualism and multilingualism.	Undecided - 6.7% (1) Agree – 43.8% (7) Strongly agree – 50% (8)

Understanding key concepts	n=16
It is important to promote the development of plurilingualism in the language classroom.	Undecided - 6.7 % (1) Agree – 62.5% (10) Strongly agree - 31.3% (5)
Language teaching should NOT aim to achieve native speaker proficiency	Disagree - 12.5% (2) Undecided - 18.8% (3) Agree – 62.5% (10) Strongly agree – 6.7% (1)
A learner's competence in a language is always "partial" and evolving.	Undecided - 6.7% (1) Agree - 68.8 % (11) Strongly agree – 25% (4)

N.B.: Due to the sample size (n=16), it was only possible to use descriptive statistics.

As can be seen from Table 14, the majority of the teachers understand and share goals of the language education with a multilingual focus. All the respondents agree that students should use the language for communicative purposes, for 73% of the teachers 'a native-speaker proficiency' is not a benchmark against which learners' language proficiency should be measured. This understanding is in line with the teacher agreement (93.8%) that a learner's competence in a language is always 'partial'. In addition, the majority of the respondents (93.8%) claims that they understand the difference between plurilingualism and multilingualism and they also acknowledge the importance of developing plurilingualism in the language classroom. At the same time, only slightly over 50% of the teachers agree that students should not learn about the language as a subject, suggesting an action-oriented approach to language teaching.

The reflection group data help us interpret the findings of the survey. In general, teachers' reflections demonstrate that they understand basic concepts that define a multilingual, plurilingual turn in education. Thus, all five teachers viewed plurilingualism as an asset with students. However, teacher 1 remarks that students' linguistic repertoire might be a hurdle in learning an additional language. She did not specify the reasons but mentioned some research report about the cases of interference in learning an additional language, which might really be the point she was making. In addition, all teachers highlight that it is crucial to develop students' repertoire in two or more languages" as student plurilingualism "provides more opportunities for students to grow and develop" (Teacher 3). Therefore, the teachers underscore that "language education should equip a learner with sufficient skills and knowledge to ensure his/her efficient communication in diverse contexts" (Teacher 2). They also explained their understanding of learners' linguistic repertoire as "the knowledge of languages students use or learnt, including students' L1". In addition, the teachers recognise that "every learner possesses an individualised and unique repertoire" (Teacher 2).

A finding of the reflection group regarding a 'native speaker standard' as a criterion against which learners' language proficiency is measured is contradictory to the finding regarding the goal of language education as presented by the respondents of the questionnaire. Four teachers of the focus group acknowledge that a 'native speaker standard' is used as a criterion in language learning when it comes to measuring grammatical and lexical accuracy, and language proficiency of pre-service teachers. At the same time, 74% of the respondents of the questionnaire report that they agree or strongly agree that the goal of language education should not be the development of native-like proficiency. In this example, there is an inconsistency between recognising the use of a native speaker standard to assess language proficiency and the belief that language education should not aim for native-like proficiency. This indicates a transitional process where teachers might be theoretically embracing plurilingualism and plurilingual assessment but are unsure about how to implement it effectively in practice (Vogt 2024).

Table 15 presents the data of the section of the questionnaire that looked into the multilingual/plurilingual practices in teaching English.

Table 15. ***The results of reported multilingual/plurilingual practices in teaching English***

Multilingual/plurilingual practices	**n=16**
In the language classroom, I ensure that my students act as social agents.	Agree – 68.8 %(11) Strongly agree – 31.3% (5)
In Ukraine, language teaching develops language students' plurilingualism (establishing the relationship between all the languages taught).	Disagree - 12.5 % (2) Undecided - 25% (4) Agree – 56.3 (9) Strongly agree – 6.3% (1)
In Ukraine, language teaching develops language students' multilingualism (teaching each language separately).	Disagree – 18.8% (3) Undecided - 31.3% (5) Agree - 50% (8)
In teaching English, I encourage my students to use other languages.	Disagree – 31.3% (5) Agree – 62.5% (10) Strongly agree – 6.3% (1)
In teaching English, I build on my students' linguistic repertoire.	Disagree - 6.3 % (1) Undecided - 6.3% (1) Agree – 68.8% (11) Strongly agree -18.8% (3)

According to the data of Table 15, promoting student plurilingualism is an important aim of language education in teaching English in Ukraine. In this regard, 62.6 % of the respondents (10 teachers) report that they develop students' plurilingualism. To support this claim, 70% (11) of the teachers allow other languages in teaching English and 87.6 % (14) of the teachers build on their student linguistic repertoire.

According to the findings of the reflection logs, all respondents admit that students' linguistic repertoire is a valuable resource that can be used to assist their students' progress in learning an additional language. However, in the opinion of the focus group, teaching practices in Ukraine foster additive multilingualism. This finding contradicts the collected data of the questionnaire where the majority of the respondents (60%) agree that in the English classroom teaching practices promote learner plurilingualism rather than multilingualism. The contradiction between the opinion expressed in the reflection log and the questionnaire responses regarding teaching practices in Ukraine can be explained by the conflicting perceptions and transitory nature of language teachers in this educational context. On the one hand, teachers may not necessarily expect to contribute to fostering students' repertoire while teaching English. On the other hand, the questionnaire responses might reflect the practical experiences of teachers who see language learning as a process where students engage with multiple languages to varying degrees rather than solely focusing on one language. When asked how teachers build on their students' repertoire, two respondents (teachers 1, 4) mentioned that they allow translanguaging and extralinguistic means of communication for the sake of meaning when it comes to teaching English to students of non-language specialisations; and teacher 2 allows L1 to translate vocabulary and explain difficult concepts.

Table 16. *The results presenting plurilingual assessments in language education*

Plurilingual assessment	n=16
Monolingual assessment fails to acknowledge complex communicative practices of plurilinguals.	Undecided - 25 % (4) Agree – 68.8% (11) Strongly agree – 6.3% (1)
In Ukraine, monolingual language assessment is a prevailing approach.	Disagree - 6.3% (1) Undecided - 31.3% (5) Agree – 50% (8) Strongly agree – 12.5% (2)
In Ukraine, plurilingual assessment is coherent with teaching English.	Strongly disagree – 6.3% (1) Disagree – 18.8% (3) Undecided – 18.8% (3) Agree – 56.3 % (9)
In Ukraine, approaches to language assessment should be reconceptualised from the standpoint of plurilingualism.	Undecided – 37.5% (6) Agree – 50% (8) Strongly agree – 12.5% (2)
Plurilingual assessment tasks should be used to assess my students' proficiency in English.	Undecided - 12.5% (2) Agree - 87.5% (13)
I understand what language assessment tasks should be used to engage all linguistic resources of my students.	Agree – 75% (12) Strongly agree - 25% (4)
I understand how to select and adapt the descriptors from the CEFR/CV relevant to a language task.	Agree – 81.3 % (13) Strongly agree – 18/8% (3)

Table 16 shows that 60% of the teachers agree that monolingual language assessment is prevailing in Ukraine; and 31.3% have not decided whether assessment in Ukraine targets at one language only. The most significant finding of the questionnaire is that above 56.3% (9) of the respondents find that plurilingual assessment actually reflects teaching practices in the English classroom. This might be the reason for 62.5% of the teachers in the study to agree that language assessment should be reconceptualised from the standpoint of plurilingualism. In this respect, the study reveals a significant change in the teachers' perspective, suggesting a departure from approaches that may have centred on monolingual standards or assessments. This shift reflects a growing recognition among the respondents of the need to adapt assessment practices to better reflect the multilingual realities of contemporary language learning contexts.

Another finding of the survey indicates that the workshop equipped the teachers with strategies for creating an assessment task which can engage students' plurilingual resources, and empowered them with an understanding of how to select and adapt the CEFR/CV descriptors relevant for a plurilingual language task. In a transitory situation like the one the teachers seem to find themselves in, it is vital to provide language teachers with relevant strategies and practices to bring the change they seem to embrace theoretically. The data of the reflection log demonstrate that the teachers unanimously believe that plurilingual assessment is a requirement of the foreign language classroom today. However, there is a danger that these responses may be influenced by social desirability bias, where teachers might provide answers they believe are expected or valued by the researchers or their peers (Lavidas et al. 2022). This bias can distort the results, potentially misrepresenting the true opinions and attitudes of the teachers involved. Despite this concern, the arguments that the teachers offer to advocate for the reconsideration of the approaches to language assessment are compelling. Teacher 1 highlights that

otherwise assessment does not reflect how communication really happens. Teachers 2 and 5 underscore that language assessment should reflect the modern requirements of an authentic communicative task. Teacher 3 highlights that assessment tasks should be developed to measure learners' ability to use their other languages in diverse multilingual situations. According to Teacher 4, "monolingual language assessment fails to acknowledge complex communicative practices of plurilinguals and their ability to draw on their diverse linguistic repertoire". Therefore, the respondents clearly see affordances of plurilingual assessment and seem to embrace it despite the fact that assessment practices in Ukraine have not followed suit.

5 Discussion

Ukraine is a multilingual country, with the majority of the population speaking several languages to different levels of proficiency (Myhre et al. 2021; Osidak and Natsiuk 2024). In this context, "tests should match actual language practices and multilinguals use resources from their whole linguistic repertoire" (Gorter and Cenoz 2017: 243). The teachers in the study report that other languages including Ukrainian have often been present in a variety of teaching activities (explaining difficult concepts, defining vocabulary, translation tasks, translanguaging). However, the analysis of the samples of assessment tasks demonstrates that the prevailing approach to test construction is monolingual. Given that the data involves only Ukrainian and English, it might be more accurately described as a bilingual rather than truly plurilingual approach. This limited inclusion of languages may not fully capture the diverse linguistic repertoires of plurilingual learners, thus restricting the potential to assess and support their plurilingual competencies comprehensively.

Another finding of the sample test analysis regards the validity of using written translation of creative texts tasks to assess the language proficiency of pre-service teachers. As it is noted in the CEFR/CV (COE 2020: 44), "translating a written text in writing is a formal process related to the activities of professional translators". The analysis of teacher assessment practices has revealed that translation as a common assessment task in the Ukrainian context and translation from Ukrainian into English is often used to assess knowledge of vocabulary use. In this respect, Flognfeldt et al. (2020) underscore that the foregrounding of translation as a plurilingual assessment task is indicative (again) of the transitory, ambivalent phase of plurilingual assessment in which teachers have a positive attitude towards plurilingual assessment but lack the means to implement it in their classrooms. This finding is also in line with the conclusions made in other studies (e.g., Simensen 2007; Studer and Kelly 2023). The analysis of the submitted plurilingual assessment tasks demonstrates that to promote plurilingualism in language education and create assessment tasks that will provide conditions for learners to engage with their other languages, several strategies were successfully employed: using crosslinguistic mediation of a text in writing or speech; engaging multilingual resources; applying language awareness and pluricultural awareness. Most of the participants of the project included only the state language in order to modify monolingual tasks to the plurilingual context, indicating a monolingual paradigm for assessment (Dendrinos 2019). In this respect, Flognfeldt et al. (2020) report that allowing students to build on their linguistic resources in a language classroom may be a challenge for educators and managing more than one language can be seen as a problem for teachers. As a result, the inclusion of only the state language by most participants overlooks the possible linguistic diversity and the presence of other languages that participants might speak and understand. In addition, other studies report that teachers do not always consider their students' previous language knowledge to be a resource in the classroom (De Angelis 2011). Our findings reflect Flognfeldt et al.'s (2020) and Simensen's (2007) conclusion that teacher persistent adherence to one language-only (English) teaching and assessment practices may be the reflection of recently prevailing language-didactic orthodoxy.

In order to encourage teachers to bring a shift towards a plurilingual perspective in language teaching and assessment, it is important to equip them with practical tools (North and Piccardo 2023). The use of

CEFR/CV descriptors are a significant prerequisite for fostering change in assessment approaches and can be really helpful to language teachers who wish to promote a plurilingual approach by suggesting real-world oriented activities (North and Piccardo 2023). Additionally, descriptors can empower teachers to create assessments that not only measure language proficiency but also promote and recognise the diverse linguistic repertoires of their students. Based on the data of this study, the selected and adapted descriptors indicated that the teachers in this project found the CEFR/CV to be a useful instrument for designing tests and developing assessment task descriptors. The teachers also effectively customised and adapted the descriptors to suit their local context. This finding reiterates a conclusion of the study by Vogt et al. (2022) carried out among Ukrainian university teachers to investigate their familiarity and expertise with the CEFR/CV, stating an understanding of an important message of the framework that it should be tailored to and customised in local contexts. The respondents in other studies (cf. Alas and Liiv 2014) similarly valued the CEFR because it is adaptable to many language situations and local contexts. Furthermore, all 16 teachers stated that the selection of relevant descriptors for the plurilingual task helped them familiarise with the CEFR/CV (see Table 16), which might contribute to the teachers' overall improved assessment literacy (Inbar-Lourie 2017). These adapted descriptors can serve as a common reference point, facilitating consistency and coherence in plurilingual assessment practices across different educational institutions.

Both the teacher reflection responses and the findings of the questionnaire analysis evidence that the workshops urged the teachers to think about their existing assessment practices as well as teaching and learning strategies in the English classroom. According to the findings of the reflection logs, students' plurilingualism is perceived as an asset in language education and calls for teaching and assessment practices that involve all learner linguistic repertoire. This implies that students' linguistic diversity is no longer viewed as a hindrance in EFL assessment in the Ukrainian context. On the contrary, the respondents were ready to embrace it as a valuable resource. The participants in the study incorporated assessments that encourage cross-linguistic mediation, language and cultural awareness between Ukrainian and English, allowing students to draw upon their diverse linguistic resources.

Moreover, the teachers in the study realise that this necessity arises from the practical language usage requirements that are linked to the present-day linguistic diversity of society (Cutrim Schmid 2021; Stathopoulou 2020; Tai and Wong 2022). The data of the reflection log demonstrate that plurilingual activities reflect real-life and professional situations that students can find themselves in and these activities are of a great relevance to the learning goals of their courses. Consequently, data collected from both cohorts of teachers indicated the necessity to reconceptualise assessment practices so that students can draw on their plurilingual competence while completing a task. Yet, the teachers do not quite know how to implement plurilingual assessment practices. For example, in modifying assessment tasks to a plurilingual context and adapting descriptors for the language assessment, the teachers chose to centre their focus on Ukrainian and English as part of a plurilingual repertoire rather than strictly viewing it as a bilingual context. Such an approach suggests that the teachers in the study observe a bilingual rather than multilingual approach to multilingualism.

On the other hand, focusing on Ukrainian and English as a plurilingual repertoire, broadens students' understanding of language competence and enables teachers to develop descriptors that are more inclusive and reflective of the diverse linguistic realities of their students. Yet, the focus on Ukrainian and English makes the authors think that the participants may feel vulnerable if an assessment task includes languages that the test-taker does not know. By limiting assessments to English and Ukrainian, there is a risk of not fully engaging with the plurilingual reality of many learners. Consequently, such an approach may ultimately hamper the development of a truly inclusive and representative plurilingual assessment framework. Moreover, this approach fails to leverage the whole linguistic repertoire that learners bring to the classroom (COE 2020). Therefore, addressing this issue requires careful consideration of how to support and empower learners in multilingual contexts, ensuring assessments are both fair and reflective of learners' diverse linguistic capabilities.

According to Harsch and Seyfer (2020), revising existing assessment practices implies bringing changes to educational systems. Such a change cannot be managed by few teachers alone as the development and validation of the instruments is time- and resource-demanding. To effectively handle such alterations, collaborative methods that include relevant stakeholders are preferable (Harsch and Seyfer 2020). Regarding this study, developing and validating sets of criteria and test specifications that take into account the implications of a new plurilingual paradigm in an assessment task is the next step to be taken. For example, an increase in reading time is necessary in modified tasks that include an additional reading text in students' L1.

The reconceptualisation of existing assessment practices cannot be simply inserted into an existing context (Poehner and Inbar-Lourie 2020). Obviously, the change will necessitate retraining teachers and assisting them in developing their professional expertise in conducting plurilingual teaching practices in general and assessment in particular. However, this project demonstrates that the teachers' awareness and positive attitude to multilingual assessment practices has been raised and they have shown their ability to design plurilingual assessment tasks on the basis of the CEFR descriptors, which is a major prerequisite of change, providing a structured framework for teachers to implement plurilingual assessment practices effectively (North and Piccardo 2023). This experience is one of the first steps in the Ukrainian context in creating more valid tests through collaborative professional initiative with other universities.

A limitation of the study was that it analysed assessment practices contributed to the study by only three universities with different number of tests provided for analysis. Therefore, it is a small-scale study. In this light, we cannot present generalised conclusions about the assessment instruments used in Ukraine to assess students' proficiency in English. Another limitation is the sample of the participants of the questionnaire (n=16) and the reflection log (n=5). Thus, we cannot argue that findings are representative for drawing consistent conclusions, but they will still provide valuable insights.

6 Conclusion

This paper presents the outcomes of a project aimed at enhancing teachers' awareness and strategies for implementing plurilingual assessment in English language teaching. The study involved analysing assessment tests from three universities in Ukraine to evaluate the extent to which plurilingual assessment practices are implemented. This study reveals that in the realm of educational assessment, there is a notable incongruity between teachers' positive attitudes towards plurilingual assessment and the predominant adherence to monolingual assessment practices. Despite recognising the value of plurilingual assessment in providing a more authentic reflection of learners' linguistic diversity, the teachers in this study reported a lack of concrete repertoire of plurilingual assessment strategies. This disjunction underscores the tension between willingness and implementation, which means that while the teachers express readiness to embrace plurilingual assessment, they struggle with putting this intention into practice. This might explain a predominance of partly monolingual assessments in English that limits students' opportunities to showcase their plurilingual competence.

In this light, this collaborative professional development project was carried out to assist teachers' growing awareness and capacity for plurilingual assessment practices. As a part of the project, two workshops were conducted to train the teachers on plurilingual assessment aligned with the CEFR/CV framework. Results indicate teachers' readiness to adopt plurilingual assessment methods, prompting a need to reconceptualise existing monolingual approaches. The participants demonstrated proficiency in developing plurilingual assessment tasks and adapting CEFR/CV descriptors to their teaching contexts when receiving appropriate guidance. Through the adaptation of CEFR/CV descriptors to their instructional contexts, the teachers demonstrated an evolving ability to integrate plurilingual assessment strategies into their pedagogical frameworks. This collaborative approach to knowledge construction not only empowered the teachers but also positioned them as catalysts for educational

change, particularly in this transitional period. Moreover, the project underscores the instrumental role of the CEFR/CV descriptors in facilitating this shift towards plurilingual assessment. By aligning these descriptors with evolving plurilingual assessment paradigms and contextualising them within specific educational settings, the participants succeeded in effectively designing assessment tasks to employ their students' linguistic repertoire. Thus, the presented project is an evidence of the transformative impact of collaborative knowledge construction and strategic utilisation of established frameworks in navigating the transition towards plurilingual assessment practices.

7 References

Alas, Ene & Suliko Liiv. 2014. Assessment literacy of national examination interviewers and raters – Experience with the CEFR. *Eesti Rakenduslingvistika Ühingu Aastaraamat* 10. 7-22. https://doi.org/10.5128/ERYa10.01

Bisai, Samrat & Smriti Singh. 2018. Rethinking assessment – A multilingual perspective. *Language in India* 18(4). 308-319. www.languageinindia.com (accessed 9 September 2024).

Chalhoub-Deville, Micheline B. 2019. Multilingual testing constructs: theoretical foundations. *Language Assessment Quarterly* 16(4/5). 472-480. https://doi.org/10.1080/15434303.2019.1671391

Choi, Julie, Russell Cross, Larissa McLean Davies, Sue Ollerhead & Melissa Barnes. 2022. Looking towards plurilingual futures for literacy assessment. *The Australian Journal of Language and Literacy* 45. 325-340 https://doi.org/10.1007/s44020-022-00021-3

Cohen, Louis, Lawrence Manion & Keith Morrison. 2007. *Research methods in education,* 6th ed. London & New York: Routledge.

Conteh, Jean & Gabriela Meier. 2014. Introduction. In Jean Conteh & Gabriela Meier (eds.), *The multilingual turn in languages education. Opportunities and challenges,* 1-14. Bristol: Multilingual Matters.

Council of Europe (COE). 2001. *Common European framework of reference for languages: Learning, teaching, assessment (CEFR)*. COE.

Council of Europe (COE). 2020. *Common European framework of reference for languages (CEFR): Learning, teaching, assessment – Companion volume*. COE. www.coe.int/lang-cefr (accessed 9 September 2024).

Creswell, John W. 2009. *Research design: Qualitative, quantitative, and mixed methods approaches,* 3rd edn. Los Angeles, London, New Delhi & Singapore: Sage.

Christie, Agatha. n.d. *Philomel cottage*. http://celine.klinghammer.free.fr/documents%20pdf/nouvelles/Philomel%20cottage.pdf (accessed 13 January 2025).

Cummins, Jim. 2008. Teaching for transfer: Challenging the two solitudes assumptions in bilingual education. In Nancy Hornberger (ed.), *Encyclopedia of language and education,* 1528-1538. New York: Springer.

Cutrim Schmid, Euline. 2021. 'I think it's boring if you now only speak English': Enhancing learner investment in EFL learning through the use of plurilingual tasks. *Innovation in Language Learning and Teaching*. 1-16. https://doi.org/10.1080/17501229.2020.1868476

De Angelis, Gessica. 2011. Teachers' beliefs about the role of prior language knowledge in learning and how these influence teaching practices. *International Journal of Multilingualism* 8(3). 216–234. https://doi.org/10.1080/14790718.2011.560669

De Backer, Fauve, Stef Slembrouck & Piet Van Avermaet. 2020. Functional use of multilingualism in assessment: Opportunities and challenges. *Research Notes* 78. 35-43.

De Wever, Bram, Tammy Schellens, Martin Valcke & HildeVan Keer. 2006. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers and Education* 46. 6-28. http://dx.doi.org/10.1016/j.compedu.2005.04.005

Dendrinos, Bessie. 2019. Multilingual testing and assessment for plurilingual education. '*MultiTest*' *ECSPM position paper.* 1-12 http://ecspm.org/wp-content/uploads/2019/03/MultiTest.pdf (accessed 9 September 2024).

Dörnyei, Zoltan. 2007. *Research methods in Applied Linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Duarte, Joana & Ingrid Gogolin. 2013. Linguistic superdiversity in educational institutions. In Joana Duarte & Ingrid Gogolin (eds.), *Linguistic superdiversity in urban areas. Research approaches*, 1-24. Amsterdam / Philadelphia: John Benjamins Publishing Company. http://dx.doi.org/10.1075/hsld.2.02dua

Duarte, Joana & Mirjam Günther-van der Meij. 2020. Multilingual interaction in secondary education in the Netherlands: A translanguaging approach. In Christiane Kirsch & Joana Duarte (eds.), *Multilingual approaches for teaching and learning: From acknowledging to capitalising on multilingualism in European mainstream education*, 75-93. London: Routledge. https://doi.org/10.4324/9780429059674-7

Flognfeldt, Mona E., Dina Tsagari, Dragana Šurkalović & Therese Tishakov. 2020. The practice of assessing Norwegian and English language proficiency in multilingual elementary school classrooms in Norway. *Language Assessment Quarterly* 17(4). 1-22. https://doi.org/10.1080/15434303.2020.1827409

Gorter, Durk & Jasone Cenoz. 2017. Language education policy and multilingual assessment. *Language and Education* 31(3). 231-248. https://doi.org/10.1080/09500782.2016.1261892

Inbar-Lourie, Ofra. 2017. Language assessment literacy. In Elana Shohamy, Iair G. Or & Stephen May (eds.), *Language testing and assessment. Encyclopedia of language and education,* 257-271. Cham: Springer. https://doi.org/10.1007/978-3-319-02261-1_19

Hashim, Noor Aneeis & Nurahimah Mohd Yusoff. 2021. The use of reflective practice towards achieving effective English language teaching at primary schools. *International Journal of Evaluation and Research in Education* 10(1). 364-373. https://doi.org/10.11591/ijere.v10i1.20956

Harsch, Claudia & Sibylle Seyfer. 2020. Marrying achievement with proficiency – Developing and validating a local CEFR-based writing checklist. *Assessing Writing* 43. 1-15. ELSEVIER. https://doi.org/10.1016/j.asw.2019.100433

Lavidas, Konstantinos, Stamatios Papadakis, Dionysios Manesis, Anastasia Sofia Grigoriadou & Vasilis Gialamas. 2022. The effects of social desirability on students' self-reports in two social contexts: Lectures vs. lectures and lab classes. *Information* 13(10). 491. https://doi.org/10.3390/info13100491

Lopez, Alexis A., Turkan Sultan & Danielle Guzman-Orth. 2017. Assessing multilingual competence. In Elana Shohamy, Ilana Or & Stephen May (eds.), *Language testing and assessment. Encyclopedia of language and education*, 1-12. Cham: Springer. https://doi.org/10.1007/978-3-319-02261-1_6

Myhre, Marthe Handå, Olena Muradyan & Oksana Nekhaienko. 2021. Educational reform and language policy in Ukraine: Implementation in the border regions. In Aadne Aasland & Sabine Kropp, (eds.), *The accommodation of regional and ethno-cultural diversity in Ukraine. Federalism and internal conflicts*, 171-209. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-80971-3_7

North, Brian & Enrica Piccardo. 2016. *Developing illustrative descriptors of aspects of mediation for the common European framework of reference (CEFR)*. Strasbourg: COE.

North, Brian & Enrica Piccardo. 2017. Mediation and the social and linguistic integration of migrants: Updating the CEFR descriptors. In Jean-Claude Beacco, Hans-Jürgen Krumm, David Little & Philia Thalgott (eds.), *The Linguistic integration of adult migrants / L'intégration linguistique des migrants adultes: Some lessons from research / Les enseignements de la recherché*, 83-90. Berlin & Boston: Mouton de Gruyter. https://doi.org/10.1515/9783110477498-011

North, Brian & Enrica Piccardo. 2023. Plurilingualism and assessment. In Silvia Melo-Pfeifer & Christian Ollivier (eds.)*, Assessment of plurilingual competence and plurilingual learners in educational settings. Educative issues and empirical approaches*, 178-194. Routledge. https://doi.org/10.4324/9781003177197-14

Osidak, Viktoriia & Maryana Natsiuk. 2024. Towards 21st century multilingualism in Ukraine: The present landscape. In Karin Vogt & Bassey Edem Antia (eds.), *Multilingual assessment - finding the nexus?*, 259-284. Berlin: Peter Lang.

Piccardo, Enrica & Brian North. 2019. Developing new CEFR descriptor scales and expanding the existing ones: Constructs, approaches and methodologies. *Zeitschrift für Fremdsprachenforschung* 30(2). 143-161.

Poehner, Matthew E. & Ofra Inbar-Lourie (eds.). 2020. *Toward a reconceptualization of second language classroom assessment: Praxis and researcher-teacher partnership*. Cham: Springer Nature Switzerland AG. http://dx.doi.org/10.1007/978-3-030-35081-9

Schissel, Jamie L., Constant Leung, Mario Lopez-Gopar & James R. Davis. 2018. Multilingual learners in language assessment: Assessment design for linguistically diverse communities. *Language and Education* 32(2). 1-18. https://doi.org/10.1080/09500782.2018.1429463

Seed, Graham. 2020. What is plurilingualism and what does it mean for language assessment? *Research Notes* 78. 5-15.

Seed, Graham & Martine Holland. 2020. Taking account of plurilingualism in Cambridge Assessment English products and services. *Research Notes* 78. 16-25.

Simensen, Aud Marit. 2007 [1998]. *Teaching a foreign language: Principles and procedures,* 2nd edn. Bergen: Fagbokforlaget.

Sridhar, Sankaran Narayanan & Kamal Krishna Sridhar. 2018. Mixing, multilingualism, and intelligibility. *World Englishes* 37. 511-522.WILEY. https://doi.org/10.1111/weng.12340

Stathopoulou, Maria. 2020. The new CEFR descriptors for the assessment of written mediation: Exploring their applicability in a local context in an effort towards multilingual testing. *CEFR Journal - Research and Practice* 2. 40-78. https://doi.org/10.37546/JALTSIG.CEFR2-3

Studer, Patrick & Paul Kelly. 2023. From monologue to dialogue: Opening Teachers' minds to authentic interaction with students. *Journal of Multilingual Theory and Practice* 4(1). 117-138

Tai, Kevin W. H. & Chiu-Yin (Cathy) Wong. 2022. Empowering students through the construction of translanguaging space in an English as a first language classroom. *Applied Linguistics* 20. 1-53. https:// doi.org/10.1093/applin/amac069

Tsagari, Dina, Kirstin Reed & Lucilla Lopriore. 2023. Teacher beliefs and practices of language assessment in the context of English as a lingua franca (ELF): Insights from a CPD course*. Languages* 8(58). 1-23. https://doi.org/10.3390/languages8010058

Vogt, Karin. 2024. Teaching (foreign) languages in multilingual contexts – a European contextualization. In Karin Vogt, & Bassey Edem Antia (eds.), *Multilingual assessment - finding the nexus*, 63-94. Berlin: Peter Lang.

Vogt, Karin, Viktoriia Osidak & Olga Kvasova. 2022. Understanding Ukrainian university teacher's familiarity with the CEFR/CV: Establishing university teacher's profile in a local context. *Ars Linguodidacticae* 9. 4-22. https://doi.org/10.17721/2663-0303.2022.1.01

Younas, Ahtisham, Sergi Fabregues, Angela Durante, Elsa Lucia Escalante, Shahzad Inayat & Parveen Ali. 2023. Proposing the "MIRACLE" narrative framework for providing thick description in qualitative research. *International Journal of Qualitative Methods* 22. 1-13. https://doi.org/10.1177/16094069221147162

8 Biographies

Viktoriia Osidak, a PhD holder in Education, is an Associate Professor at Taras Shevchenko National University of Kyiv, Ukraine. She has authored numerous publications, with an emphasis on alternative assessment strategies to enhance L2 learning and teaching. Her research interests lie within Teaching English as a Foreign Language, multiligualism and Language Testing and Assessment.

Karin Vogt is a professor at Heidelberg University of Education. Among her research interests are the CEFR/CV, language testing and assessment, Vocationally-Oriented Language Learning and inclusive foreign language teaching. She is the author of numerous publications on different aspects related to Teaching English as a Foreign Language. Currently she is involved in research projects on multilingual and multimodal language assessment and inclusive foreign language teaching.

Maryana Natsiuk, a PhD holder in Education, Associate Professor at the Department of Methodology of Teaching Ukrainian and Foreign Languages and Literature, Educational and Scientific Institute of Philology, Taras Shevchenko National University of Kyiv, Ukraine. The scientific interests include teaching EFL, innovative methods of teaching foreign languages, teaching culture, multilingualism. She is the author of numerous publications on different aspects related to Teaching English as a Foreign Language.

Appendix A

Analysis of the test against such criteria as targeted skills, task characteristics, the language of input, the language of student performance, and the type of assessment.

University 1: Test #1: General English for communication	
Year/ programme	1st to 3rd year of Bachelor, B1+ - B2+ at CEFR Level
Skills	Speaking, writing (opinion essay), grammar (state the difference in meaning, paraphrase/ find and correct the mistake), vocabulary (translation)
Task characteristics	Monolingual, bilingual, communicative, generic, aims at an ideal native-speaker language use
Input	English, Ukrainian
Output	English
Mode of assessment	Summative
University 1: Test #2: General English for communication	
Year/ programme	1st to 4th year of Bachelor, B1+ - B2+ at CEFR Level
Skills	Interaction; mediation (Explaining data/ image); speaking; writing (an opinion essay), grammar (language focus tasks; syntactic analysis of a sentence)
Task characteristics	Monolingual, isolated, communicative, generic, aimed at an ideal native speaker language use
Input	English
Output	English
Mode of assessment	Summative
University 1: Test #3: English for Business Communication	
Year/ programme	4th year of Bachelor, B2+ - C1 at CEFR Level
Skills	Reading into speaking, reading into writing, grammar test (language focus tasks, syntactical analysis of the sentence
Task characteristics	Monolingual, integrated, communicative, generic, aims at an ideal native speaker language use
Input	English
Output	English
Mode of assessment	Standardised testing system

University 2: Test 1: English for professional purposes - consists of two parts, includes exams of winter and spring terms	
Year/ programme	1st year, Bachelor programme, B1+ - B2+ at CEFR Level
Skills	Reading into writing, speaking, interaction, translation from English into Ukrainian, grammar.
Task characteristics	Monolingual, bilingual, targets isolated skills, integrated communicative, generic, aims at ideal native speaker language use
Input	English
Output	English, Ukrainian
Mode of assessment	Summative
University 2: Test 2: English for professional purposes - consists of two parts, includes exams of winter and spring terms	
Year/ programme	4th year of Bachelor, B2+ - C1 at CEFR Level
Skills	Mediation, writing, translation from English into Ukrainian, grammar.
Task characteristics	Monolingual/ bilingual, targets isolated/ integrated skills, communicative/ discrete, generic, aims at ideal native speaker language use
Input	English
Output	English, Ukrainian
Mode of assessment	Summative
University 2: Test 3: English Communication Course	
Year/ programme	1st year Master, C1 at CEFR Level
Skills	Mediation, speaking, translation from English into Ukrainian, grammar.
Task characteristics	Monolingual/ bilingual, targets isolated/ integrated skills, communicative/ discrete, generic, aims at ideal native speaker language use
Input	English
Output	English, Ukrainian
Mode of assessment	Summative
University 2, test 4: Speak English Professionally Course	
Year/ programme	1st year, Master, C1 at CEFR Level
Skills	Mediation (relaying specific information), speaking, writing, translation from English into Ukrainian, vocabulary.
Task characteristics	Monolingual/ bilingual, targets isolated/ integrated skills, communicative/ discrete, generic, aims at ideal native speaker language use
Input	English
Output	English, Ukrainian
Mode of assessment	Summative
University 3, Test 1: A practical English Course	
Year/ programme	1st year, Bachelor, B1+-B2 at CEFR Level
Skills	Speaking/ Interaction, a language focus test (vocabulary and grammar)
Task characteristics	Monolingual targets isolated skills, communicative/discrete, generic, aims at an ideal native speaker language use
Input	English
Output	English
Mode of assessment	Ongoing, formative (a project)/ summative

Appendix B

The analysis of languages in tests to assess language proficiency of pre-service teachers

Years 1-3, Bachelor	**3 tests**	**5 tasks**	**Languages involved**
University 1: Test #1: General English for communication	Speaking	1	in one language (English)
	Writing an essay	1	in one language (English)
	Grammar	2	in one language (English)
	Vocabulary (translation)	1	in two languages (Ukrainian, English)
Years 1-4, Bachelor	**4 tests**	**6 tasks**	**Languages involved**
University 1: Test #2: General English for communication	Interaction	1	in one language (English)
	Mediation: Explaining data/ image	1	in one language (English)
	Speaking	1	in one language (English)
	Writing an essay	1	in one language (English)
	Grammar	2	in one language (English)
Year 4, Bachelor	**1 test**	**4 tasks**	**Languages involved**
University 1: Test #3: English for Business Communication	Mediation: Relaying specific information in speaking	1	in one language (English)
	Mediation: Relaying specific information in writing	1	in one language (English)
	Grammar- language focus tasks, syntactic analysis of the sentence	2	in one language (English)
Year 1, Bachelor	**2 tests**	**5 tasks**	**Languages involved**
University 2: Test 1: English for professional purposes	Translating a written text	1	In two languages (Ukrainian, English)
	Mediation: Relaying information	1	in one language (English)
	Mediating: Expressing a personal response	1	in one language (English)
	Interaction	1	in one language (English)
	Grammar	1	in one language (English)
Year 4, Bachelor	**1 test**	**5 tasks**	**Languages involved**
University 2: Test : English for professional purposes	Translating a written text	1	In two languages (Ukrainian, English)
	Mediation: Relaying specific information	1	in one language (English)
	Mediation: Expressing a personal response	1	in one language (English)
	Interaction	1	in one language (English)
	Grammar	1	in one language (English)

Year 1, Master	1 test	5 tasks	Languages involved
University 2: Test 3: English Communication Course	Translating a written text	1	In two languages (Ukrainian, English)
	Mediation: Relaying specific information	1	in one language (English)
	Mediation: Expressing a personal response	1	in one language (English)
	Grammar	2	in one language (English)
Year 1, Master	**1 test**	**6 tasks**	**involved Languages**
University 2, test 4: Speak English Professionally Course	Translating a written text	1	In two languages (Ukrainian into English)
	Mediation: Relaying specific information	1	in one language (English)
	Speaking	1	in one language (English)
	Mediation: Relaying specific information in writing	1	In two languages (Ukrainian into English)
	Grammar	1	in one language (English)
	Vocabulary	1	in one language (English)
Year 1, Master	**1 test**	**1 tasks**	**Languages involved**
University 3, Test 1: A practical English Course	Speaking: A project with a built-in progression	1	in one language (English)

Appendix C

Selecting and adapting relevant descriptors from the CEFR/CV to the assessment task: Prepare a collaborative group project that incorporates the topics, vocabulary, and grammar structures learned throughout the course. In addition to the course material, utilise podcasts, interviews, videos, and blogs in other languages that you know (including L1) related to the course topics. The project should be presented in English and can take the form of a video, performance, or presentation. Present a reference list of the sources used.

Criteria	Level	Descriptors from the CEFR/CV	Adaptation of the descriptors
Overall mediation	B2+	Can establish a supportive environment for sharing ideas and facilitate discussion of delicate issues, showing appreciation of different perspectives, encouraging people to explore issues and adjusting sensitively the way they express things. Can build on others' ideas, making suggestions for ways forward. Can convey the main content of well-structured but long and propositionally complex texts on subjects within their fields of professional, academic and personal interest, clarifying the opinions and purposes of speakers/signers (COE 2020: 92).	Can establish a supportive environment for sharing ideas and facilitate discussion of delicate issues, which can arise in the process of collating multilingual sources, showing appreciation of different perspectives, encouraging people to explore issues and adjusting sensitively the way they express things. Can build on others' ideas, making suggestions for ways forward. Can convey the main content of well-structured but long and propositionally complex texts on subjects within the given task.
Relaying specific information in speech or sign	B2+	Can relay (in Language B) which presentations given (in Language A) at a conference, or which articles in a book (in Language A) are particularly relevant for a specific purpose (COE 2020: 94).	Can relay (in English) which presentations, video, articles given (in English, Ukrainian and other languages) are particularly relevant for a specific purpose
Processing text	B2+	Can summarise clearly in well-structured language (in Language B) the information and arguments contained in complex texts (in Language A) on a wide range of subjects related to their fields of interest and specialisation (COE 2020: 99).	Can summarise clearly in well-structured language (in English) the information and arguments contained in complex texts (in English, Ukrainian and other languages) on a wide range of subjects related to their task.
Processing text	B2	Can synthesise and report (in Language B) information and arguments from a number of sources (in Language A) (COE 2020: 100).	Can synthesise and report (in English) information and arguments from a number of sources (in English, Ukrainian and other languages).

Criteria	Level	Descriptors from the CEFR/CV	Adaptation of the descriptors
Facilitating collaborative interaction with peers	B2+	Can act as rapporteur in a group discussion, noting ideas and decisions, discussing these with the group and later giving a summary of the group's view(s) in a plenary (COE 2020: 110).	No adjustment
Collaborating to construct meaning	B2+	Can contribute to collaborative decision making and problem solving, expressing and co-developing ideas, explaining details and making suggestions for future action (COE 2020: 110).	No adjustment
Strategies to simplify a text: Amplifying a dense text	B2+	Can make concepts on subjects in their fields of interest more accessible by giving concrete examples, recapitulating step by step and repeating the main points (COE 2020: 122).	Can make concepts on subjects in their fields of interest more accessible by giving concrete examples, recapitulating step by step in order to perform the task
Building on pluricultural repertoire	B2	Can, in an intercultural encounter, recognise that what one normally takes for granted in a particular situation is not necessarily shared by others, and can react and express themselves appropriately (COE 2020: 125).	Can, as an intercultural encounter, recognise that what one normally takes for granted in Ukrainian culture is not necessarily shared by others.
Plurilingual comprehen-sion	B2	Can use their knowledge of contrasting genre conventions and textual patterns in languages in their plurilingual repertoire in order to support comprehension (COE 2020: 126).	No adjustment
Building on plurilingual repertoire	B2	Can alternate between languages in their plurilingual repertoire in order to communicate specialised information and issues on a subject in their field of interest to different interlocutors (COE 2020: 128).	Can alternate between Ukrainian and English in their plurilingual repertoire in order to communicate specialised information and issues of their task

Appendix D
Teacher reflection log

Understanding the basic CEFR/CV related concepts

1. What is the difference between multilingualism and plurilingualism?
2. What is your learners' language repertoire?
3. What is the difference between plurilingual and multilingual students?
4. Why is it important to promote the development of plurilingualism in the language classroom and assessment?
5. What is a 'partial' language competence?

Plurilingual practices in the English classroom

6. Do you consider your learners' repertoire?
7. Do you use "a native speaker's standard" as a criterion in language learning?
8. What are the objectives of language education in Ukraine?
9. Does language teaching in Ukraine include developing language students' plurilingualism?

Plurilingual assessment practices in the language classroom

10. Why should approaches to language assessment be reconsidered?
11. Would plurilingual assessment tasks be compatible with your existing assessment framework?

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Awareness-Raising in Training Student Teachers to Rate Written-Performances in line with the CEFR

Jana Bérešová, Trnava University, Slovakia

https://doi.org/10.37546/JALTSIG.CEFR7-3

Language teaching in Slovakia is based on the concepts presented in the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) and CEFR Companion Volume (CEFR/CV). Despite efforts to align the learning outcomes in primary and secondary education to CEFR proficiency levels in the national curricula and significant changes in approaches to language teaching and learning in previous years, teachers' familiarity with particular level descriptors is unbalanced. Their marking of students' written performances does not explicitly reflect the different proficiency levels. To change this situation in the country, intensive training of student teachers became necessary.

Assessing written performances requires systematic training based on marking criteria related to a particular reference level. To prepare future teachers of English to be able to distinguish between particular proficiency levels, several activities were designed to ensure that student teachers acquire a detailed knowledge of a targeted set of descriptors. Among other documents, a written assessment grid (Appendix 4, CEFR/CV) was introduced to sensitise student teachers to the need to familiarise themselves with the descriptors and apply them, specifying what is appropriate to expect at different levels of achievement.

A new academic course on assessment and testing of language competence has been introduced in conjunction with a pre-service training course. The aim is to ensure that student teachers receive the necessary training to apply marking criteria when assessing written and oral performances. This study explores the approach adopted using a written assessment grid from the CEFR Companion Volume. Particular activities and data that were collected and analysed during the course of the present study are furthermore presented.

Keywords: written performances, rating, pre-service teacher training, raters' judgements, reference descriptors

1 The impact of the CEFR on language education in Slovakia

The first provisional version of the Common European Framework (CEF), as it was initially called, in 1996 and 1998, later published as *A Common European Framework of Reference for Languages: Learning, Teaching Assessment* (COE 2001), significantly influenced language education in Slovakia. This was a time when the educational system itself in Slovakia was in transition due to political changes, and language learning – a subject area that had been underestimated for decades – required a substantial shift in focus from one focused on learning language systems to an approach that focuses on the use of languages for real-life purposes.

The concepts that underpin the CEFR were immediately introduced in the school-leaving examination reform for foreign languages in 1997 when particular descriptors for reception skills were applied in test construction. Such impact is evident in the compilation of new standards (Štátny pedagogický ústav 1999) and new curricula (Bérešová et al. 2002) for foreign languages taught in primary and secondary education, such as English, French, German, Italian, Russian and Spanish, which are currently the

languages of the school-leaving examination. Curriculum writers were inspired by 'can do' statements, which in turn were incorporated into learning outcomes. Alongside officially claimed requirements (Bérešová et al. 2002; Štátny pedagogický ústav 2016), they are now considered the alpha and omega for item writers in the construction of tests for different reference levels.

1.1 The impact of the CEFR on the Slovak school-leaving examinations in languages

The reform of the school-leaving examination was initiated by language teachers who sought to change the assessment of learners' foreign language proficiency. The objective was to establish a valid and reliable measurement. Previously, final examinations conducted in schools lacked objectivity and evidence of reliability and validity in the measurement tools. This led to the implementation of high-stakes examinations, which provide stakeholders with valid and reliable data on secondary school leavers' language competence, transparently displayed on the website of the National Institute for Education and Youth (https://www2.nucem.sk/sk/merania/narodne-merania/maturita).

After a three-year piloting process of English tests, the Ministry of Education officially recognised the monitoring process and accepted the necessity of introducing an external part of language school-leaving examinations administered by the ministry-governed testing institution. This process ended in 2004-2005, and since then, an external part of high-stakes language examinations has been administered every year, except for two years of the pandemic situation in the country (2020 and 2021). However, from the very beginning, English has been assessed at two levels, called lower and higher, based on the students' selection. In 2008-2009, following a large number of interventions into the system of testing foreign languages, the Ministry of Education officially recognised three levels of completing language education for secondary school leavers: B1 for students studying at secondary technical schools, B2 for school-leavers from secondary grammar schools and C1 for students studying at bilingual schools or bilingual sections of secondary schools. All the requirements based on the CEFR 'can do' descriptors and officially recognised standards are available in the Catalogues of Requirements for each level (Štátny pedagogický ústav 2016) and school-leaving examination specifications, modified regularly and adapted to specific situations if needed.

Initially, language teachers were enthusiastic about the changes and getting objective data about their students' achievements. The government was, therefore, urged to adopt more objective methods for the assessment of writing and speaking skills as well. However, the costs of hiring teachers to assess papers or oral performances, administrative costs and employing statisticians hindered progress towards consistency between external measures of listening comprehension, reading comprehension, language in use and the measurement of two productive skills.

1.2 The current status of the issue

After twenty years, the situation is entirely different. The enthusiastic teachers who volunteered in the activities carried out as part of the piloting process were replaced by a new generation of teachers facing new challenges, such as earning money to survive in current economic conditions and coping with a lack of EFL teachers in the state sector. Due to the overload of classes (26-32 hours per week), teachers of English rely on published documents or coursebooks printed in the UK and labelled with proficiency levels and do not commonly consult either an English version of the CEFR or its Slovak translations (Spoločný európsky referenčný rámce pre jazyky: učenie sa, vyučovanie a hodnotenie 2006, a revised version in 2017). Teachers are familiar with the common reference levels of the CEFR. However, they may not have a comprehensive understanding of the model of language-related competences and language use. This task demands a thorough comprehension of and specific reference to descriptive examples that are pertinent to the CEFR levels that English teachers work with.

When the provisional document *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): a Manual* was published in 2009, the

Slovak Republic started the process of aligning language tests with the CEFR, publishing several articles on the processes, such as familiarisation, standardisation, benchmarking, and standard setting. As the setting of cut scores requires the involvement of several parties in a multi-stage, judgemental process, test developers, item writers, policymakers (Berk 1986), teacher trainers and language teachers were invited to participate in several workshops to discuss a linking process. Consultants from the European Centre for Modern Languages were also invited for a workshop to train test developers and item writers to design tests and construct items in line with the CEFR.

Despite efforts to adopt approaches to assessment that are in line with the CEFR, the Ministry of Education remains reluctant to do any research. It supports the previously set cut scores at 33% of test achievement (listening, language in use and reading) and 25% in testing writing. If a test taker achieves less, they fail. If they achieve more than cut scores, test takers are seen as learners of a respective level. Despite criticism in this regard, decisions have so far remained unchanged. As a result of this ministerial approach, test developers and item writers have stopped working on the task of aligning language examinations with the CEFR.

On the other hand, one needs to consider the new generation of teachers who have entered the profession and are not familiar with the process of alignment of examinations with the CEFR initiated in previous years. Although changes in language teaching and testing are significantly influenced by the CEFR (COE 2001) and CEFR/CV (COE 2020), language teachers are the driving force in terms of the practical implementation and application of CEFR philosophies and related concepts. Language teachers need to be thoroughly informed about the aims of language education and to comprehend consistencies and discrepancies between practical findings and theoretical postulations. Only then can the required changes be operationalised.

In-service teacher training used to be organised through a well-developed system that catered for life-long education provided in eight regional centres, situated in each regional capital city. Currently, workshops for language teachers are not commonly organised since language teaching is not the focus of in-service teacher training. It can be concluded from consultancy experience that there is a gap between knowledgeable and more experienced teachers who received regular training some years ago and new teachers graduating from several universities with methodology courses having different aims that are not necessarily aligned with the CEFR perspective.

In addition, not all the methodology courses at Faculties of Arts or Faculties of Education focus specifically on the CEFR. Future language teachers are consequently not aware of concepts and approaches emphasised in the CEFR and CEFR/CV. In workshops, language teachers usually claim that they are familiar with particular reference levels; however, it soon becomes apparent that they have never gone through the process of familiarisation during which participants undergoing training grasp the notions of particular descriptors. Becoming familiar with the wording of particular descriptors is crucial as it enables both students and their teachers to see what students can do at a particular level of proficiency, matching students' performances against relevant descriptors. However, the most challenging aspect is that the matching actual performances against described competences should be justified and enough evidence needs to be provided.

2 Methodology

Following an analysis of problems Slovak teachers of English face (Bérešová 2019, Bérešová 2020), Trnava University introduced a new academic course related to assessing and testing learners' language competence in English. In their master-degree programme, student teachers are presented with fundamental considerations in language testing (Bachman 1990; Bachman and Palmer 1996; Council of Europe 2011), test construction (Alderson et al. 1996; Weir 1993), statistical analyses (Bachman 2006), and other test-related topics as the basis for seminars during which student teachers experience particular aspects, applying theory in practice.

During the practicum at primary and secondary schools, student teachers are exposed to much assessment-based input, which is strongly influenced by traditional testing. Their supervisors use formative tests for summative purposes and spoken and written performances are not assessed against any marking criteria in the classroom context. Student teachers thus report their experiences in the ELT seminars and expect to be provided with evidence on how testing and assessment can be qualitatively improved.

2.1 Study questions

The academic course on assessment and testing of communicative competence had 18 enrolled Master's degree students. To respond to their expectations for measurement of communicative competence in line with the CEFR, it was essential to plan some activities that would enable them to become aware of the processes necessary for the implementation of CEFR 'can do' statements into rating learners' written performances. The focused question about the relationship between training and being able to rate learners' written performances in line with the CEFR was the following:

RQ: Does one-semester training affect the student teachers' ability to rate learners' written performances in line with the CEFR?

2.2 Participants

To obtain proper data, it was essential to design intensive training for 18 student teachers (15 females, 3 males) who were required to participate in all activities related to the procedures necessary for their mutual understanding of specific reference levels and illustrative descriptors used in various scales. They were obliged to participate continuously for three months. In case of their face-to-face absence, they joined the group online.

2.3 Research methods

This study employed quantitative research methods, complemented by qualitative analyses of two written performances. To collect data, student teachers' ratings based on the CEFR written grid (CEFR/CV 2020) and those based on the rating scale for written performances (Appendix B) used in the country in line with the Catalogue of Requirements for Level B2 (Štátny pedagogický ústav 2016) were calculated and summarised in the tables. Quantitative methods provided quantified background data to contextualise the presented study. The data provided a basis for a detailed analysis of particular descriptors in both documents. Qualitative methods helped in data interpretation.

Then, student teachers were required to assess two performances written by B2 learners, who, as secondary grammar school leavers, were officially expected to apply for a B2 test. Qualitative analysis of both performances based on the judgement of written performances against qualitative aspects acted as a source of intuiting, which was then tested by quantitative measurement (student teacher's ratings). The ratings were compared with the official scores of the expert teachers who were in charge of assessing learners' performances according to the officially recognised marking criteria (Appendix B).

2.4 Materials

The materials used in both student teachers' assignments were the same: the written assessment grid (CEFR/CV 2020), the rating scale (Appendix B), and a B2 task (Appendix A). For a qualitative analysis, three online tools, such as the *English Vocabulary Profile* (https://www.englishprofile.org/wordlists), the *English Grammar Profile* (https://www.englishprofile.org/english-grammar-profile) and *Text Inspector* (https://textinspector.com/), were useful for obtaining detailed data about the quality of two secondary school-leavers' written performances.

The written performances that student teachers were expected to match to the reference levels, and consequently assess, were part of the school-leaving examination that is a high-stakes examination, externally run by the testing centre. In 2022, the topic of this externally-administered written task, internally assessed by designated teachers, was Transport and Travelling (Appendix A).

2.5 The period of awareness raising

This study examines particular procedures and results in the academic course initiated in March 2023 and concluded in a session that culminated in a debate when problematic issues were presented and discussed in May 2023. Two hours per 10 weeks enabled students to become aware of fundamentals (4 hours), become familiar with a variety of scales and illustrative descriptors (4 hours), become aware of the coverage of the high-stakes examination, the official specification of writing and the officially recognised marking criteria (2 hours), to be presented with standardisation procedures (2 hours), experiencing the process of rating productive skills against the qualitative aspects (2 hours), judging the first performance (2 hours), judging the second performance (2 hours) and comparison of data and final discussion (2 hours).

2.5.1 Fundamentals and introduction to awareness-raising stages

Language test development and examination is a challenge not only for test developers and item writers but also for language teachers. As mentioned above, language testing is a minor topic of methodological courses. Language teachers strive to prepare their students for being tested with limited knowledge related to language test construction and evaluation, which is a complex area and requires theoretical background as well as practical experience.

The CEFR raised many questions concerning the testing of language competence, which was later reflected in the *Manual for Language Test Development and Examining* (COE 2011). These developments have largely contributed to the resources that make up the Council of Europe's toolkit, the users of which need to familiarise themselves to be able to make effective use of the CEFR in their own contexts. The CEFR proposes a general model of language use and language learning. To operationalise this model in language testing, two aspects of authenticity (situational and interactional) must be considered while constructing test items and tasks (COE 2011). Language testing can be viewed from different perspectives, but fundamental considerations that underlie the practical development and use of language tests, proposed by Bachman (1990), significantly influenced the CEFR model of language use. Due to a growing need of the users that decided to follow the CEFR paradigm shift and tended to change the nature of language assessment by aligning their language tests and examinations to the CEFR, *Relating the language examinations and tests to the Common European Framework of Reference: Learning, teaching, assessment – Manual* (COE, 2009) was regarded as encouragement in their endeavours to situate their national language examinations with the CEFR perspective.

Five-interrelated stages described thoroughly in the *Relating the language examinations and tests to the Common European Framework of Reference: Learning, teaching, assessment – Manual* (COE, 2009) emphasised the necessity of being consistent in demonstrating the validity of the claims made about the relationship between language examinations and the levels of the CEFR. The validation of the claim requires both theoretical and empirical evidence. The linking process presupposes standard setting referring to content standards and performance standards that are both defined in the CEFR in the form of level descriptors.

To make students familiar with the interrelated stages of situating tests in relation to the CEFR, student teachers were invited to be actively engaged in the activities and all stages recommended in the *Relating the language examinations and tests to the Common European Framework of Reference: Learning, teaching, assessment – Manual* (COE 2009). Focusing on the project, only the activities concerning written production will be mentioned in this study.

2.5.2 The stage of familiarisation

The main aim of the familiarisation stage was to help student teachers become aware of CEFR 'can do' descriptors related to written production. The scales selected for writing were the three most relevant ones, namely those for *Overall written production*, *Creative writing* and *Reports and essays* (COE 2020). Each scale contained a randomly changed order of the levels without being indicated. Student teachers were expected to indicate a level, underline key words in the descriptors provided and discuss it in contrast with other levels. During this activity, it was possible to recognise the influence of their previous experience of learning English during their secondary-school studies when their teachers focused on the accuracy of grammar and vocabulary rather than text types, established conventions of the genre, the complexity of discourse at higher proficiency levels, grammar accuracy and vocabulary range. It was essential to emphasise the difference between B1 and B2 in argumentation while writing essays, as B2 writers are expected to give some reasons in support of or against a particular point of view, and B1 can write one-sidedly (COE 2001). To support student teachers' awareness of linguistic competences, the scales for *Vocabulary range, Vocabulary control and Grammatical accuracy* (COE 2020) were integrated into training.

2.5.3 The stage of specification

The specification stage included student teachers' familiarisation with the official specifications of three school-leaving examination levels and awareness of what learners are expected to perform in writing. The transparently displayed specification for testing writing at level B2 raised discussion on the number of tasks, and student teachers agreed upon a minimum of two tasks to get more evidence about language learners' competencies, as proposed by Weigle (2002). However, a school-leaving examination contains only one task due to several reasons, such as the length of an examination, the burden placed on language teachers who are to mark their students' written performances due to a lack of funding for external raters as well as the status of this type of testing – externally assigned and internally marked. Table 1 displays the official specification obligatory for item writers and assessors for level B2.

Table 1. ***Specification for Writing at Level B2***

Aim	To measure learners' ability to write independently, genre-based, stylistically and grammatically adequately, clearly, comprehensibly and at an appropriate level stated in the Catalogue of Requirements.
Time	60 minutes
Number of tasks	One task, thematically consistent with the topics presented in the Catalogue of Requirements.
Tested skills and subskills	Task achievement, the use of adequate linguistic structures, a range of vocabulary, composition and stylistics, paragraphing and orthography.
Task	A structured task based on the written input. The task is structured into 3-5 points.
Range/scope	200-220 words (minimal number of words: 120)
Rating criteria	Officially recognised marking criteria

Task design is viewed as the most challenging stage in testing writing. Although it is reasonable to state that being given a choice of prompts to write on may be preferred by students, writing on different topics can potentially make the results less reliable (Weigle 2002). Therefore, in Slovakia, students are

given structured tasks on a particular topic, embedded in one of the CEFR domains in which social life is organised (COE 2001) and differentiated on the basis of proficiency levels. Based on Harsch and Rupp (2011), the school-leaving examination can be viewed as a level-specific examination aimed at assessing and reporting school-leavers' proficiency with a focus on one proficiency level. In the project, the focus was only on one reference level (B2), and the intention was to discuss learners' performances in terms of the CEFR descriptors. It is implied that besides the relationship between the targeted proficiency level and the task characteristic, the relationship between the proficiency level and the rating scale level needs to be transparent (Harsch and Rupp 2011).

When being given CEFR descriptors for level B2, student teachers estimated that the task given to students to test their language competence in English enabled them to provide enough evidence on being able to write clear, detailed texts, synthesising and evaluating information and arguments. In line with theoretical approaches, the task needs to provide an opportunity for language learners to show their range (Tardieu et al. 2010; COE 2001). The number of arguments to address also makes the task more challenging and an array of processing and reasoning required to solve the task (Harsch and Rupp 2011). The latter was confirmed by student teachers while assessing the difficulty of the task (Appendix A). The descriptors related to the B2 level describing general linguistic range, vocabulary range, grammatical accuracy and vocabulary control, as well as orthographic control, were consistent with the expected complex language requirements.

Each task should be consistent with marking criteria. At this stage, student teachers were presented with the officially recognised marking criteria in the form of an analytic scale. According to Harsch and Rupp (2011), there are not many studies on the effect of holistic or analytic criteria on the variability of level-specific ratings. However, their study aimed at presenting the data on task difficulty, rating criteria difficulty, and other aspects that influence rating variability, such as raters and learners' abilities (Harsch and Rupp 2011). Conversely, Rea-Dickins and Germaine (1998) imply that analytic-marking schemes have the advantage of providing diagnostic detail of use as learner's performance is described at a range of different areas. Moreover, analytic scoring (sometimes called multi-trait) enables raters to evaluate different aspects of performance separately (Weigle 2012).

The school-leaving exam rating scale (Appendix B) covers six bands (from 0 to 5) focusing on four aspects of written production: task achievement (content), organisation (genre, coherence and cohesion, stylistic quality), grammar (syntactic variability and complex grammatical structures) and vocabulary (range, variability and appropriacy). Regarding the first aspect, student teachers discussed the CEFR descriptors related to thematic development and propositional precision, namely developing a clear argument, clearly signalling the difference between fact and opinion or passing on detailed information reliably (COE 2020) and compared them with the descriptions in the examination rating scale. The same process was followed, using scales related to coherence and cohesion mentioned in the second marking criteria. For the purposes of grammar and vocabulary areas, a variety of scales (general linguistic, vocabulary range, grammatical accuracy, vocabulary control and orthographic control) encompassed in the 'can do' statements describing linguistic competence (COE 2020) were then judged and juxtaposed with the descriptors in the rating scale. The latter were estimated to be consistent with level B2 descriptors taken from the CEFR. However, during the discussion of problematic areas in relation to a partial inconsistency in the judgements of student teachers, the following point emerged: the wording of the rating scale descriptors for each band seemed to be challenging and student teachers agreed upon a necessity of intensive training to ensure consistency of marking.

2.5.4 The stage of standardisation

Standardisation is seen as a process of consensus building concerning what learners can do at a given level and whether that corresponds to the level claimed by the resource (British Council, UKALTA, EALTA and ALTE 2022). As it is based on arriving at a common understanding of what a language learner can

do at a particular CEFR level, the calibrated performances seemed relevant for training student teachers to become aware of existing exemplary resources.

One of the valuable materials containing illustrative samples of production, already aligned to the CEFR, are the DVDs, which enable us to gain a clear understanding of the relevant CEFR level. Although the DVDs provide performances of oral production, they are a good demonstration of the performance quality at the required level. The calibration was based on the criteria grid that contains similar qualitative aspects, namely range, accuracy and coherence (COE 2001), which were considered as consistent with 'can do' statements referring to any production performance. Student teachers worked individually, matching the selected performances with CEFR descriptors to become aware of what language learners are expected to do at level B2, namely in terms of grammar and vocabulary.

2.5.5 The stage of standard setting

When describing standard setting, there are many methods for setting cut scores that should be based on a generally accepted methodology and reflect the judgement of qualified people (Zieky and Perie 2006). Standard setting can be viewed as 'a blend of judgements, psychometrics, and practicality' (Hambleton and Pitoniak 2006: 235); however, judgements are commonly considered the cornerstone on which cut scores are based. Regarding test items or examinees, the two main approaches adopted are test-centred methods and examinee-centred methods. Using judges' estimations, it is possible to recognise inconsistencies since their judgements are influenced by their experience (e.g., thinking about students they had taught), and they employ different standards when placing students into performance categories (Van Nijlen and Janssen 2008; Engelhard 2009; Bérešová 2017).

Reckase (2009) summarises the standard-setting process, which is usually required by the 'agency,' and claims that the final numerical score needs to be consistent with test design and content, elaborated description and policy definition of a standard. However, different methods or different implementations of the same method used in standard setting may not provide results that are of equal quality.

As regards testing writing, holistic judgements on work samples seem to be relevant. It is the Body of Work method (COE 2009), which allocates the student's performance to one of the predefined levels for which panellists are to set the standard. This method is commonly set on two rounds; if more are needed, a third round can be added. The scores of the students' performances are not known by the panellists, and their judgements are converted into cut scores, using logistic regression (Noijons et al. 2011).

North (2014) claims that the first method that was proposed to situate results on a test to several levels was the Carroll method, based on the use of "real data from teacher assessments and piloting it against the real test scores of the same group of learners" (North 2014: 216). This standard setting focuses on a correlation between the two sets of results for the same learner.

Standard setting is embedded in the empirical process of gathering quantitative evidence to make appropriate cut-score scales (British Council et al. 2022). This applies to any standard-setting methodology. Once employing test-centred methods, judges estimate at what reference level a test taker can be expected to respond correctly to a set of items. However, when testing writing, the concern is that examinee-centred methods sometimes referred to as empirical-judgemental methods (Berk 1986), where someone who knows test takers provides a holistic assessment of the CEFR levels are applied. However, the analytic judgement method, mentioned in *The CEFR Alignment Handbook* (British Council et al. 2022), is based on reviewing actual performances on the writing test. The performances are expected to be scored by "trained raters using the scoring scales developed for the test in question" (British Council et al. 2022: 57).

This stage of the course was more theoretical than practical. Student teachers were provided with the above-mentioned theoretical approaches to standard setting to allow them to understand the complexity and importance of the standard-setting process. It deserves a great deal of attention and a professional approach. The reason for not going through the standard-setting process was that two

main preconditions were unmet, such as experienced assessors and an insufficient number of samples. However, this stage aimed to make student teachers aware of how important this process is once test takers need proof of their language competence. Standard setting, officially documented, enables stakeholders to judge the quality of the assessment.

3 Results

A group of 18 student teachers experienced an intensive awareness-raising training to analyse and assess test tasks and performances in relation to the CEFR levels. As part of this study, two real-life pieces of work were the subject of two different scoring procedures. In one case, the rating scale (Appendix B) was applied first and only then, the performance was linked to the target level; in the other case, the CEFR assessment grid (COE 2020) was applied before the rating scale.

Due to a lack of time, limited by the hours of the academic course, it was possible to provide student teachers only with two written performances. Since the processes of judging those performances differed, both are described separately to clearly recognise particular steps and problems that occurred while working on the assignments.

The ratings of 18 students are presented in the tables to clearly show the student teachers' judgments. Their final ratings of the learners' written performance are compared with those of the officially appointed assessors. The data obtained from the online tools are presented in the tables to analyse inconsistencies in the assessments.

The school that enabled student teachers to assess real performances disclosed only information that could be provided with respect for confidentiality. Other learners' scores, namely those achieved in the external part of the B2 examination and the Speaking test, were added to their written performances to get a complete picture of the learners' abilities. Student teachers were not informed of these achievements or the scores the performances received from the officially appointed teachers during the process of their rating.

3.1 Student teachers' ratings of the first written performance

The student teachers were given the marking criteria (Appendix B) and one school leaver's performance. Based on the marking criteria, each aspect can be given 5 points as the best performance and 0 as the lowest performance. Using the marking criteria and linking the learner's performance with the task formed the first round of judgements (Table 2), revealing that student teachers were most consistent while rating learner's ability to organise their text, meeting a majority of the characteristics of the genre, linking all the ideas mostly logically, and using appropriate connectors. Other aspects were judged in two different bands. The most significant difference emerged while rating the grammatical competence of the learner as one group of raters focused on correctness. In contrast, the rest focused on a range of grammatical patterns and the use of complex language expected at level B2.

Table 2. ***Student teachers' ratings in the first round—the first learner's performance***

Points	Task achievement	Organisation	Grammar and spelling	Vocabulary
4	13	18	8	3
3	5	-	10	15

In the second round, student teachers were asked to work in smaller groups of 3 or 4. Consulting all the previously given CEFR scales, they had to present their estimations justifying their judgements. After the second round and a long discussion, student teachers agreed upon the final estimate for all four

qualitative aspects of written performance as 4+4+3+3, converted into 70% of successful performance. During their justification, it was revealed that while discussing the performance, they had not linked the rating scale performance descriptions and CEFR descriptors with the performance but had used their previous learning experience or had been influenced by their practicum. Therefore, a different approach directly linked to the CEFR descriptors was used in their assessment of learners' performances.

Student teachers were asked to use a written assessment criteria grid (COE 2020) and estimate CEFR level of performance. Their judgments were distributed amongst three reference levels, although B1+ is not officially worded in the grid (Table 3). Several recognised that the performance does not fully match any of the officially formulated performances and estimated intuitively that the level between B1 and B2 might be B1+. As the aspect of the overall performance at B2 includes descriptors, such as *can write clear, detailed official and semi-official texts on a variety of subjects* or *can make a distinction between formal and informal language,* the aspect of accuracy entails *showing a relatively high degree of grammatical control and not making errors that cause misunderstandings*, and a learner could write *an essay, which develops an argument* (COE 2020), most student teachers estimated that the learner could perform at level B2 in these three areas.

Table 3. ***Student teachers' judgements based on the written assessment criteria grid—the first learner's performance***

Levels	Overall	Range	Coherence	Accuracy	Argument
B2	13	8	4	10	10
B1+	2	10	2	-	-
B1	3	-	12	8	8

Based on the yes/no judgement round, it can be concluded that the overall performance of the first student was estimated at level B2; however, while judging range and coherence, most student teachers claimed that the performance did not match B2 descriptors. Therefore, student teachers were addressed to discuss their judgements precisely and to provide supportive arguments for their choice. In their pre-service teacher training academic course, student teachers became aware of the labelled words both in the Cambridge Learner's Dictionary (McIntosh 2013), gained from the production of test-takers, while the Oxford Learner's Dictionary (https://www.oxfordlearnersdictionaries.com/) word labelling is based on the words used in the English coursebooks published in the Oxford University Press. During evidence claims, student teachers consulted words labelled in the English Vocabulary Profile and found that only five words used in the learner's paper are labelled B2, such as *affect*, *firstly*, *means*, *pollute* and *harmful*.

In this case, it was decided to use the Text Inspector system (https://textinspector.com/) to measure the quality of the learner's performance (Table 4). At the vocabulary level, the word list types revealed that the learner's performance expected at level B2 was represented by lower-level types of words. The total number of analysed tokens was 245, of which nine were unlisted. Analysing the number of words labelled by CEFR levels, a large number of used words was more relevant for A levels users as B levels were represented only by 12% out of all the words used in the paper. This supported the student teachers' uncertainty when they were asked to judge the criterion concerning the range of language used to express opinions. As far as grammar is concerned, the sentences were in present and past tenses, once the learner used *to be going to* and *will.* There were mistakes when he/she wanted to use more advanced patterns.

Table 4. *Words and their labels taken from the Text Inspector system*

CEFR levels	A1	A2	B1	B2
Words/%	90 (64.75%)	22 (15.83%)	13 (9.35%)	5 (3.60%)
Tokens/%	185 (75.51%)	26 (10.61%)	19 (7.76%)	5 (2.04%)

The second round was less variable as student teachers focused on other descriptors related to general linguistic range and could not match the first student's performance with B2 level descriptors, such as *developing arguments without much conspicuous searching for words, using some complex sentence forms to do so* or *can vary formulation to avoid frequent repetition, shows a relatively high degree of grammatical control, can use a variety of linking words efficiently to mark clearly the relationships between ideas* (Tardieu et al. 2010). The most useful information appeared on the right side of the scale related to pragmatic competence, in which B2 descriptors give a clear approach to what a B2 user is expected to do in the language, such as *can highlight the most important aspects of a topic, can employ the rules that concern going from the general to details, can deliver all of the contents and components that are expected for the text concerned* (COE 2020). The findings showed that 44% of raters estimated the performance to be B1, reasoning that the use of language is lower than what learners can do at level B2. However, 66% were consistent in estimating the performance as B2, providing a lot of evidence, matching the first learner's performance with exact descriptors of the B2 reference level.

To conclude the rating of the first learner's written performance, it is necessary to disclose the official rating of the paper. As mentioned above, every paper was assessed by two assessors. After synchronising both judgements, the test taker was given 3+3+2+2 = 10 points (50%). The assessors' notes in the paper showed that his/her initial score (9 points = 45%) after the first round of assessment was crossed out and replaced by a new score after the second round. It can be inferred that it may arise from different factors, for example, due to his/her better achievements in an external part of the school-leaving examination, labelled as B2. The officially administered measurements of his/her listening, language in use and reading reached 78.3% (72 percentile), and his/her spoken performance was marked as 1, the best mark in the marking system. However, once CEFR reference levels are implied, his/her written performance slightly contradicts his/her ability to perform receptively. It is significantly in contrast with the achievements in another productive skill (speaking). Since there is no evidence of the learner's spoken performance during the discussion with student teachers, it seemed to be reasonable to disregard the mark for this performance. The achievements in the external part of the school-leaving examination in B2 English and the written performance proved that learner's language competence is at B2.

3.2 Student teachers' ratings of the second performance

The second paper was judged differently. The first process of estimation was based on the use of the written assessment criteria grid (CoE 2020). Having experienced the first paper estimation, student teachers started to read a 269-token long text without focusing on the task, though matching the performance against the criteria described in the grid. In the first round of their initial judgement, their estimation arrived at two levels – C1 and B2. Once the decision is C1, it can be concluded that the learner can achieve B2 (Table 5).

Table 5. ***Student teachers' judgements based on the written assessment criteria grid – the second learner's performance***

	Overall	Range	Coherence	Accuracy	Argument
C1	-	5	5	-	-
B2	18	13	13	18	18

Student teachers were positively impressed by a lengthy text containing much information produced by the second learner. The language constructions, such as -ing and -ed participles, the use of different tenses and the use of advanced cohesive devices influenced their estimation. However, when they were given the marking criteria (Appendix B) and the task (Appendix A), their first estimation resulted in the rating presented in Table 6. Using the marking criteria, the performance appeared to be weaker than that of the first student. The weakest aspect was task achievement, as student teachers could recognise memorised parts of the text that did not match the task.

Table 6. ***Student teachers' ratings in the first round – the second learner's performance***

Points	Task achievement	Organisation	Grammar and spelling	Vocabulary
4	2	-	-	-
3	6	18	18	8
2	10	-	-	10

During the discussion, student teachers admitted that after the first reading without focusing on the task, the second performance seemed to be written by a good user of English, providing a lot of information in a more advanced language. However, it did not match the bullet points that were clearly stated in the task, enabling markers to be more objective, not being biased by learner's ability to produce a lot of language related to the topic, but not matching the task.

The Text Inspector data proved that the learner used two C1 words; however, one of them (commuting) was used in the rubrics, and the second (sector) has its Slovak form with a letter k. In addition, it can be inferred that the learner used four B2 words, such as *causing*, *secondly*, *traffic jam* and *decade*. In contrast to the previous learner's performance, the number of unlisted words was 17 types representing 19 tokens and, due to their misspelling, such as 'almoust,' 'busses,' and 'enourmous,' they were not included in the labelling process. Comparing both text inspector data, it can be concluded that the second performance contained more A2 and B1 words, but the text produced seems to be based on the text from the coursebook the learner memorised while preparing for an oral examination to achieve a good mark in speaking.

Based on the English Grammar Profile, the structures used in the second paper were labelled B1. The same conclusion can be made, using the Slovak Catalogue of Requirements for B2, in which exact structures are mentioned and categorised. When student teachers participated in the second round, their estimations became more consistent, and their assessments were unequivocally 2+3+3+2, which finally meant 10 points. It can be concluded that the performance was given a 50-percent success rate following the transfer of points into percentages.

As mentioned above, a pragmatic competence scale includes descriptors related to the scales of coherence and thematic development. While the second learner could structure the text logically, maintaining a clear development, the text he/she produced was not based on the bullet points of the task, but on the topic, in essence. However, the learner could make links between different parts of the text and construct the text by applying rules that involve moving from the general to the detailed.

The official raters seem to have been biased as well. While in the first round of the judgement, their decision was 2+2+2+2 = 8 points (40%), after the second round, the learner achieved 3+3+3+3 = 12 (60%). The achievements in the external school-leaving examination were 78.3% (72 percentile), and speaking was marked with the highest mark (1). Based on the use of the CEFR, it can be concluded that the performance of the second learner can be labelled B2, although some doubts arose during the rating process.

3.3 Conclusive remarks on the rating process

In training student teachers to rate B2 written performances in English, the analytic judgement method procedures were followed. Although two different approaches were used: the rating scale (Appendix B) as the first and the written assessment criteria grid (CEFRCV 2020) as the second in the first learner's rating and vice versa in the second learner's rating did not significantly influence student teachers' judgements. The failure to adhere to the structured task resulted in student teachers being initially impressed by the second learner's performance. However, they later realised that the learner had not produced appropriate content as it was related to the topic rather than the task. The officially appointed assessors gave the second learner lower scores as the main criterion in the rating scale (Appendix B) is task achievement. The decisions made by designated assessors did not have an impact on student teachers, as they were not informed about the availability of official scores. The precise information was obtained afterwards.

According to the official regulations in the country, language learners can achieve only one score point higher in other criteria than in the first criterion, ensuring that memorised text used inappropriately to complete the task cannot enable learners to pass. If task achievement is scored as 0, all other aspects of the assessment are to be marked as 0.

The descriptions of the summative profile were assessed through binary judgements as to whether the learner's performance demonstrates the required characteristics or not, as suggested by Brindley (2001). Due to qualitative analysis and careful reading of the CEFR, its descriptors and CEFR-related documents, both performances judged holistically matched reference level B2.

4 Discussion and conclusions

Each government has its language policy, considering the educational background and history of language education in a respective country. As the CEFR is descriptive, it enables policymakers, curriculum and test developers, teachers, and language users to use it so that each country or group of people affected can benefit from it.

This study aims to highlight the importance of awareness-raising training and the factors that can influence marking and raters' decisions. These factors include the task, marking scale, raters, and their training. The CEFR provides scales with descriptors that enable raters to match learners' or test takers' performances against specific reference levels. A shared understanding of specific levels can reduce inconsistencies between raters. Encouraging raters to justify their ratings and exposing them to other raters' opinions during discussions after each round can also help. It is essential to provide intensive training to ensure a common understanding of the reference scales, their level descriptors, and illustrative samples.

This study presents an overview of the stages required for student teachers to feel confident in demonstrating that their assessments are in line with the respective reference level. The most challenging aspect was assessing learners' written performance and justifying that the scores given were within the reference level. Once the ratings were agreed upon, student teachers could compare them with the scores of officially appointed teachers. This confirmed to the student teachers that the areas they had identified as problematic were similar to those identified by other evaluators. However, the student

teachers' approach to addressing the issues was consistent with CEFR-based materials. Although the number of student teachers included in this study was limited, the findings support further research in both pre-service and in-service teacher training to obtain more data on assessors' ability to combine particular CEFR level descriptors with the marking criteria to achieve consistency in rating language learners' performances at a particular level. Consequently, it is also essential to incorporate a more significant number of samples of language learners' written performances. This would enable gathering sufficient evidence to validate the claim about the relationship between learners' ability to use English in written production and a specific CEFR level.

The idea behind this study was to emphasise the significance of training the users of the CEFR. Achieving a common understanding of reference levels and descriptors is a rigorous task. The training materials used were all aligned with the CEFR, including a task, marking criteria, the written assessment grid, the English Vocabulary Profile, the English Grammar Profile, Text Inspector, and the wording used in discussions. The keywords used throughout were taken from the CEFR descriptors. The student teachers were actively involved in the process as they recognised the opportunity to experience detailed reading of the CEFR descriptors presented in different scales. They were able to discover the many possibilities of applying the CEFR in various contexts. This knowledge should be reflected in the development of classroom materials.

Based on the study results, it is recommended that a re-evaluation of teaching practices should be undertaken. Involving language teachers in the implementation of the CEFR-based marking criteria and in the construction of level-descriptor matching tasks can aid learners in performing better by familiarising them with the requirements for a particular proficiency level. Improving one's understanding of CEFR levels and their descriptors can positively impact the development of classroom materials and, subsequently, enhance the learning process.

5. References

Alderson, J. Charles, Caroline Clapham & Dianne Wall. 1996. *Language test construction and evaluation.* Oxford: OUP.

Bachman, Lyle F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, Lyle F. 2005. *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, Lyle F. & Adrian S. Palmer. 1996. *Language testing in practice.* Oxford: Oxford University Press.

Bax, Stephen. 2012. https://textinspector.com/ (accessed May 2023).

Bérešová, Jana. 2017. *The impact of the CEFR on language examinations in local contexts.* Frankfurt am Main: Peter Lang.

Bérešová, Jana. 2019. The importance of objectivity in assessing writing skills. *INTED Proceedings,* Valencia: IATED, 4376-4380.

Bérešová, Jana. 2020. An analysis of B2 tests in English in the Slovak educational context. *EDULEARN Proceedings, 12th International Conference on Education and New Learning Technologies*. Valencia: IATED, 4833-4838.

Bérešová, Jana, Marta Hosszúová & Marta Macková. 2002. *Učebné osnovy pre štvorročné gymnáziá.* [The national curriculum for four-year secondary grammar schools]. Bratislava: Štátny pedagogický ústav.

Bérešová, Jana, Marta Hosszúová, Marta Macková & Edita Mareková. 2002. Katalóg cieľových požiadaviek na vedomosti a zručnosti maturantov z anglického jazyka – úroveň B1 a B2. [The catalogue of English language requirements – levels B1 and B2]. Bratislava: Štátny pedagogický ústav.

Berk, Ronald A. 1986. A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research,* 56(1), 137-172. https://doi.org/10.2307/1170289.

Brindley, Geoff. 2001. Outcomes-based assessment in practice: Some examples and emerging insights. *Language Testing* 18(4): 393-407. https://doi.org/10.1177/026553220101800405.

British Council, UKALTA, EALTA & ALTE 2022. *Aligning language education with the CEFR: Handbook.* https://www.ealta.eu.org/documents/resources/ CEFR%20alignment %20handbook.pdf (accessed 25 April 2023).

Council of Europe. 2001. *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 (CEFR): A manual*. Strasbourg: Council of Europe.

Council of Europe. 2011. *Manual for language test development and examining.* Strasbourg: Council of Europe.

Council of Europe. 2020. *Common European framework of reference for Languages: Learning, Teaching, Assessment. Companion Volume.* Strasbourg: Council of Europe.

Engelhard, George, Jr. 2009. Monitoring raters in performance assessments. In Gerald Tindal & Thomas M. Haladyna (eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* 261-287. Mahwah, NJ: Erlbaum. https://www.taylorfrancis.com/chapters/mono/10.4324/9781410605115-17/monitoring-raters-performance-assessments-george-engelhard-jr-gerald-tindal-thomas-haladyna (accessed 8 May 2023).

English Grammar Profile. https://www.englishprofile.org/wordlists (accessed May 2023).

English Vocabulary Profile. https://www.englishprofile.org/wordlists (accessed May 2023).

Hambleton, Rronald K. & Mary J. Pitoniak. 2006. Setting performance standards. In Robert L. Brennan (ed.), *Educational Measurement* (4th edn.) 433-470. Westport, CT: American Council on Education/ Praeger.

Harsch, Claudia & Alexander André Rupp. 2011. Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quaterly* 8(1). 1-33. https://doi.org/10.1080/15434303.2010.535575

McIntosh, Colin. 2013. *Cambridge Advanced Learner's Dictionary.* Cambridge: Cambridge University Press.

Národný inštitút vzdelávania a mládeže [The National Institute for Education and Youth]. 2023. Merania. Maturita. [Measurements. The school-leaving examination.] https://www2.nucem.sk/sk/merania/narodne-merania/maturita (accessed April-May 2023).

Noijons, José, Jana Bérešová, Gill Breton & Gábor Szabó (eds.). 2011. *Relating language examinations to the Common European framework of reference for languages: Learning, teaching, assessment (CEFR). Highlights from the Manual.* Graz: European Centre for Modern Languages / Strasbourg: Council of Europe.

North, Brian. 2014. *The CEFR in Practice.* Cambridge: Cambridge University Press.

Oxford Learner's Dictionaries. https://www.oxfordlearnersdictionaries.com/ (accessed April-May 2023).

Rea-Dickins, Pauline & Kevin Germaine. 1998. Evaluation. Oxford: Oxford University Press.

Reckase, Mark D. 2009. Standard setting theory and practice: Issues and difficulties. In Neus Figueras & José Noijons (eds.). *Linking to the CEFR levels: Research perspectives.* Arnhem: Cito, EALTA. https://www.ealta.eu.org/documents/resources/Research_Colloquium

_report.pdf (accessed 28 April 2023).

Štátny pedagogický ústav [The National Institute for Pedagogy]. 1999. Vzdelávací štandard pre gymnázium – začiatočníci [Educational standards for secondary grammar schools – beginners]. Bratislava: Štátny pedagogický ústav.

Štátny pedagogický ústav [The National Institute for Pedagogy]. 2006. Spoločný európsky referenčný rámec pre jazyky: učenie sa, vyučovanie, hodnotenie. [The Common European Framework of Reference for Languages: Learning, teaching, assessment]. Bratislava: Štátny pedagogický ústav.

Štátny pedagogický ústav [The National Institute for Pedagogy]. 2016. Cieľové požiadavky na vedomosti a zručnosti maturantov z anglického jazyka – úrovne B1 a B2. [The catalogue of English language requirements – levels B1 and B2]. https://www.statpedu.sk/files/articles/nove_dokumenty/cielove-poziadavky-pre-mat-skusky/anglicky-jazyk_b1b2.pdf (accessed 10 May 2023).

Štátny pedagogický ústav [The National Institute for Pedagogy]. 2017. Spoločný európsky referenčný rámec pre jazyky: učenie sa, vyučovanie, hodnotenie. [The Common European Framework of Reference for Languages: Learning, teaching, assessment]. Bratislava: Štátny pedagogický ústav.

Tardieu, Claire, Raili Hildén, Magda Lehmann, & Monique Reichert. 2010. *The CEF-estim grid.* Graz: European Centre for Modern Languages / Strasbourg: Council of Europe.

Van Nijlen, Daniel & Rianne Janssen. 2008. Modeling judgments in the Angoff and contrasting-groups method of standard setting. *Journal of Educational Measurement* 45(1). 45-63.

Weigle, Sara Cushing. 2002. *Assessing Writing.* Cambridge: Cambridge University Press.

Weigle, Sara Cushing. 2012. *Assessing Writing*. In Christine Coombe, Peter Davidson, Barry O'Sullivan & Stephen Stoynoff (eds.), *The Cambridge Guide to Second Language Assessment.* Cambridge: Cambridge University Press.

Zieky, Michael, & Marianne Perie. 2006. *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service. http://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf (accessed 12 May 2023).

6 Biography

Jana Bérešová is a professor at the Faculty of Education of Trnava University, Slovakia, where she is involved in pre-service teacher training. In her academic courses on methodology, she introduces the CEFR and its concepts, which significantly influence language teaching, and shares her expertise with student teachers. She introduced a new academic course on testing communicative competence, which helps prepare student teachers to become aware of basic considerations in language testing and to better understand CEFR levels and descriptive 'can do' statements and their role in assessment procedures. Twenty years ago, she initiated a reform of the country's school-leaving exams, focusing on testing language use in line with the CEFR. Still an item writer, she contributes to improving the effectiveness of the examinations developed and administered in the country. She has written several monographs on language teaching and testing, articles on various aspects of language education and has presented her work at international conferences in Europe and beyond. As a member of the RELANG team initiative offered by the ECML in Graz, Austria, she contributes to changes in language education in several European countries, focusing mainly on aligning curricula, language tests and examinations with the CEFR, and encouraging local examination providers to ensure quality in high-stakes testing.

Appendix A

In your English class, you have to write an opinion essay entitled 'Transport and Travelling in My Life' (200-220 words). Follow these points:

- the influence of the transport and travelling on the quality of your everyday life,
- your positive contribution to the environment – your choice of travelling/commuting,
- an unforgettable experience from travelling by any means of transport.

Appendix B

Table 7. *An analytic rating scale to assess writing*

	Relevance and adequacy of the content	Discourse (genre, organisation)	Grammar	Vocabulary
5	The content is totally relevant to the task. All the points of the task are thoroughly and evenly elaborated. Main ideas are consistently developed.	The text meets all the characteristics of the genre. All the ideas are clearly and logically linked. Wording and cohesive devices are effectively used.	Grammar structures are used accurately and appropriately for the task to a large extent. The text is characterised by adequate syntactical variability and complex grammatical structures. Grammatical and syntactical errors occur sporadically.	Vocabulary is rich and relevant to the topic. The text is characterised by rich lexical variability, appropriate collocations and idioms.
4	The content is relevant to the task. All the points are adequately but not evenly elaborated. In general, the main ideas are developed.	The text has a majority of characteristics of the genre. Logical linking of ideas prevails. Wording and cohesive devices are appropriately used.	Language structures are used accurately and appropriately for the task to a considerable extent. The text is characterised by certain syntactic variability and complex grammatical structures. Grammatical and syntactical errors occur to a limited extent.	Vocabulary is rich and prevailingly relevant to the topic. The text is characterised by proper lexical variability and correctly used collocations and idioms.
3	The content is almost relevant to the task. One point is not adequately elaborated. The main ideas are sufficiently developed, but not all are relevant.	The text does not have a majority of characteristics of the genre. Ideas are not always sufficiently linked. Wording and cohesive devices are sufficiently used.	To a certain extent, language structures are used accurately and appropriately for the task. The text is characterised by limited syntactical variability and complex grammatical structures to a small extent. Grammatical and syntactical errors occur more frequently.	Vocabulary is appropriately rich and relevant to the topic. The text is characterised by minor flaws in using collocations and idioms.

	Relevance and adequacy of the content	Discourse (genre, organisation)	Grammar	Vocabulary
2	The content is relevant to the task to a limited extent. Two points are not adequately elaborated. The main ideas are not sufficiently comprehensible.	The text has the characteristics of the genre to a limited extent. The flow of the ideas is, for the most part, not linked logically. Wording and cohesive devices are used in a limited way.	Language structures are used accurately and appropriately to the task to a lesser extent. The text is characterised by sporadic syntactical variability, and simple structures prevail. Some grammatical and syntactical errors interfere with comprehension of the text.	Vocabulary is simple and not always relevant to the topic. The repetition of the same words characterises the text. The incorrect use of some words interferes with the comprehension of the text to a certain extent.
1	The content is minimally relevant to the task. The points are not sufficiently elaborated. The main ideas are not comprehensible to a large extent.	The text has the characteristics of the genre to a minimal extent. The flow of the ideas is not logically linked, which causes incomprehension. Wording and cohesive devices are inadequately used.	Language structures are often used inadequately and inappropriately for the text. The text is characterised by minimal syntactic variability and basic grammatical structures. Grammatical and syntactic errors interfere with comprehension of a certain part of the text.	Vocabulary is simple and relevant to the topic to a limited extent. The text is characterised by frequent repetition of the same vocabulary. The incorrect use of the words often causes misunderstanding.
0	The content is not relevant to the task. Points are elaborated by irrelevant ideas. The main ideas are not comprehensible.	The text does not have the characteristics of the genre. The flow of ideas is chaotic and illogical. Wording and cohesive devices are not used, which causes incomprehension.	Language structures are used prevailingly inappropriately and inaccurately to the task. The text is not characterised by syntactic variability and contains basic grammatical constructions. Grammatical and syntactical errors prevent understanding of the major part of the text.	Vocabulary is very simple, prevailingly irrelevant to the topic. The text is characterised by limited vocabulary. The incorrect use of vocabulary prevents understanding to a large extent.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Assessing writing proficiency in a Saudi Arabian university: Comparing students, tutors, and raters' assessment using selected CEFR scales

Ebtesam Abdulhaleem, King Salman Global Academy for Arabic Language, Riyadh, KSA

Claudia Harsch, Universität Bremen, Germany

https://doi.org/10.37546/JALTSIG.CEFR7-4

This study explores the writing proficiency levels of Saudi Arabian medical track students after completing a one-year Preparatory Year Programme (PYP), as well as the applicability of the Common European Framework of Reference for Languages (CEFR) in assessing their proficiency. The standardized writing exam administered at the end of the PYP revealed a ceiling effect, with the majority of students achieving high scores, despite the fact that the PYP teaches English at three different levels (beginner, intermediate, advanced). To obtain a more nuanced understanding of students' writing skills, alternative assessment methods were explored using selected CEFR scales, including self-assessment, tutor assessment, and assessment by raters recruited from the UK (experts in using CEFR scales). The study aimed to determine if these CEFR-based assessments can reliably differentiate among the three PYP levels, and if the CEFR scales are practical and applicable in this context. The findings show that the CEFR-based scores from all three assessor groups can reliably separate students according to their PYP level. The results highlight that the CEFR can serve as a valuable tool for understanding students' writing proficiency, even in non-European settings. This study encourages further exploration in the use of CEFR scales to assess proficiency levels.

Keywords: Writing proficiency, Preparatory Year Programme (PYP), Common European Framework of Reference for Languages (CEFR), self-assessment, tutors' assessment, raters, proficiency levels, CEFR scales

1 Introduction and background

Writing in English is a skill that many Saudi students find exceptionally challenging (McMullen 2009; Shukri 2014). This is true even among highly proficient students (Shukri 2014). To address this, Saudi Arabia has implemented Preparatory Year Programmes (PYPs) aimed at enhancing students' English skills during their initial year at university. These programmes aim to equip students with the necessary proficiency to navigate the English-medium academic environments of various colleges they will join after completing the PYP (Ebad 2014).

At the beginning of the PYP, students are grouped into three proficiency levels (elementary, intermediate or advanced) based on their test scores on the Oxford Placement Test (OPT) (OUP 2001), which evaluates students' listening and reading skills, along with grammar and vocabulary knowledge. However, the OPT does not assess written or oral skills, so proficiency in those areas remains unidentified prior to the PYP.

The OPT is scored between 0 and 100. Students scoring 0–45 are placed in the elementary level, those scoring 46–85 in the intermediate level, and those scoring above 85 in the advanced level.

At the end of the PYP, all students, regardless of the level they attend, take the same standardized proficiency exam, which includes a writing component. The exam only requires students to write a minimum of 120 words in 60 minutes on an easy, general descriptive topic about their daily routine at the university (see Appendix 1 for two performance examples, the exam itself cannot be published). It was designed based on a very low benchmark (roughly equivalent to CEFR level A2). The results of the exam revealed a ceiling effect, with scores concentrating at the upper end of the grading scale: 73% of all students achieved the highest score (10/10), regardless of the PYP level they had attended. The median and interquartile range (IQR) scores were 9.6 (9.2, 10), 10 (9.6, 10), and 10 (10, 10) for students starting the PYP at the elementary, intermediate, and advanced levels, respectively. While these high scores might indicate progress due to instruction during the PYP, or suggest that the exam was not adequately challenging, or had an insufficiently discriminating marking scheme, they do not effectively differentiate between students or accurately describe their proficiency according to an internationally recognized framework such as the CEFR. Consequently, determining the students' 'true' proficiency levels by the end of the programme proved to be challenging.

Methods that could be used to differentiate between students' levels may include assessments by the students themselves, by their teachers, or by independent raters. All methods may have advantages and disadvantages.

Self-assessment may be unreliable, since low-proficiency students tend to overestimate their proficiency (Babaii et al. 2016; Blue 1988; Leach 2012; Ünaldı 2016;). This has been described as the "metacognitive deficits" of the "Dunning-Kruger effect", i.e., it takes a certain level of competency to be able to assess one's own proficiency (Kruger and Dunning 1999). Self-assessment may also be inaccurate due to students' lack of experience in this approach (Babaii et al. 2016; Engelhardt and Pfingsthorn 2013).

Conversely, higher proficiency students may underestimate their own proficiency level (Kruger and Dunning 1999; Hodges et al. 2001; Lejk and Wyvill 2001; Tejeiro et al. 2012), possibly due to students being over-modest (Kun 2016). At the highest proficiency, researchers described more similarities between the students' and their teachers' assessment and therefore considered self-assessment as more accurate at higher-proficient levels (Kun 2016; Ünaldı 2016; Sahragard and Mallahi 2014).

As noted by Paris and Winograd (1990), familiarisation with and instruction in this approach can improve the accuracy and reliability of self-assessment. One way to determine the accuracy of self-assessment is to compare it with other methods, such as tutors' judgments or other test scores (Abdulhaleem and Harsch 2018; Ashton 2014; Babaii et al. 2016; Boud 1991), although high correlations between self-assessment and other measures of performance are unlikely (Dunning et al. 2004). For example, Falchikov and Boud (1989), in their meta-analysis of studies comparing self-assessment with teachers' marks, reported an average correlation of r=0.39. Correlation between self-assessment and students' - 'actual performance' (e.g., scores in a test) was very low (r=0.21) (Falchikov and Boud 1989).

In a similar way, teacher assessment may show comparably low correlations with scores allocated by external raters or with scores from standardized tests. Fleckenstein et al. (2018) found a correlation of r=0.41 between tutor assessments and test scores, noting that teachers overestimated students' levels compared to their actual performance in an achievement test. This overestimation was similarly evident in Bérešová's (2011) study, where teachers tended to overestimate students' vocabulary, grammar and language use compared with actual test results.

The CEFR proficiency framework has been employed to assess students' proficiency levels within Europe and beyond (e.g., Atai and Shoja 2011; Dragemark Oscarson 2009; Ünaldı 2016). Moreover, the CEFR is already used at the PYP curriculum, mainly to articulate the programme's objectives and to choose textbooks for each of the PYP levels. The principal reasons for the use of the CEFR in our study were the fact that it is already used in the PYP, the CEFR's design, and its role as a common metalanguage.

The CEFR "can be presented and exploited in a number of different formats, in varying degrees of detail" (Council of Europe [CoE] 2001: 36). The descriptors correspond well with the communicative teaching paradigm (Green 2012). Descriptors can "specify learning objectives in terms of situation, activities, functions and notions" (Green 2012: 21); and each descriptor "is worded in positive terms, even for lower levels" (North 2014: 55). The CEFR is used to "foster mutual understanding" across different users (Tannenbaum and Wylie 2005: 41); as a reference tool for identifying learners' needs prior to designing the curriculum (Little 2007); and as "a point of departure" (North 2014) to start the reflection, analysis and discussion of potential university standards and admission criteria (Harsch 2018). There are 53 CEFR scales representing different language skills and these must "be used selectively" (North 2014: 11) to suit the context in which they are employed.

2 Aim of the study

Although several studies have been conducted on Saudi students' writing skills in general (Aljumah 2012; Alkubaidi 2014; Hellmann 2013; McMullen 2009; McMullen 2014; Obeid 2017; Oraif 2016), to our knowledge, none has investigated the writing proficiency of Saudi medical track (MT) students in relation to the CEFR. The main objective of the study was therefore to obtain a more nuanced understanding of students' writing proficiency than the current exam upon completion of the three levels of the PYP-MT allows. Moreover, by comparing CEFR-based assessment from the perspectives of students and their tutors, we set out to explore the applicability of the CEFR in the Saudi Arabian PYP context, where the CEFR is not commonly used and where participants have not yet been thoroughly familiarised with this framework. Hence, students and their teachers assessed the end-of-year performances (from the standardised exam) against a CEFR-based assessment grid that contained selected CEFR writing scales. To triangulate the findings from within the PYP context, the same student performances were also assessed by external raters familiar with the CEFR, using the Writing Grid from the manual for relating language examinations to the CEFR (CoE 2009). We aimed to explore new ways of assessments that could reliably differentiate students (thus avoiding the aforementioned ceiling effect), while simultaneously benchmarking the three PYP levels against an internationally recognised framework (i.e., the CEFR). Hence, it was important to understand the extent to which scores given by students, their tutors and independent raters were comparable and correlated with each other.

Research questions

The study addresses the following research questions:

RQ1: Can students' self-assessment, tutors' assessment, and raters' assessment (using selected CEFR scales) reliably differentiate students' writing proficiency among the three PYP levels?

RQ2: To what extent are the scores from the three assessor groups comparable, taking the three PYP levels into account?

3 Methods

3.1 Overall design

The study takes a cross-sectional quantitative design. Three assessor groups assessed the same students' writing proficiency: students, their teachers and external raters. Students and their teachers assessed students' general writing proficiency, using similar assessment grids based on selected CEFR scales. Raters assessed the students' performances elicited by the end-of-year exam, using the CEFR grid from the Manual. The resulting scores from these three groups were quantitatively analyzed. The extent to which each group of assessors was able to discriminate reliably between the three PYP levels (RQ1) was analyzed using ANOVA and comparisons of means between levels, with pairwise comparisons

between each pair of levels (elementary vs. intermediate vs. advanced). The scores obtained from all three assessor groups (RQ2) were compared between pairs (students vs. tutors vs. raters) using ANOVA and independent t-tests.

3.2 Participants

The study targeted female students in the PYP-MT, as they are being prepared to enter various medical and healthcare-related colleges such as the Colleges of Medicine, Pharmacy, Dentistry, Nursing, and Applied Medical Studies. The entire female cohort of students in PYP-MT (N=640) in 2016 was invited to participate, resulting in a total of n=517 participants across the three PYP levels (elementary, intermediate, and advanced). Of the participants, 90% were Saudi and 10% were non-Saudi, aged 18–19 years.

Furthermore, all PYP tutors (N=24) teaching English to the students in the PYP-MT were also offered the opportunity to participate, with a total of n=19 tutors accepting the invitation. All participating tutors were only teaching one level (either elementary, intermediate or advanced) when the data were collected, to try to reduce any 'norm-orientation' (comparison of a student with students in other levels) during data collection, although some tutors had previous teaching experience in teaching the other levels. The study analysis included a total of n=517 students whose general proficiency was assessed by both themselves and their tutors.

To examine students' and tutors' scores in relation to external measures, seven raters from two language institutes in the UK, who were experienced with writing assessment in higher education, familiar with the CEFR framework and experienced with using CEFR-based rating scales for rating second language texts, were invited to participate and accepted. They assessed the end-of-year performances by a subsample of 105 of the 517 students who participated in this study.

3.3 Ethics

Ethical permission was granted by the University of Warwick regarding the application, instruments and data collection (as part of a PhD study). Official permission was also given from the Dean of the PYP and the PYP research committee to collect data on the women's campus and to analyze the students' final exam written texts. All participants were fully informed about the aims of the research and the consequences of their participation (Punch 2005), and that it was possible to withdraw from the study at any time during or after participation; they were also given the chance to ask any questions regarding the study. All participants received an information sheet about the study, including all relevant contact information and a consent form to be signed. Both were translated into Arabic to ensure full comprehension.

3.4 Instruments

Due to administrative constraints, we were unable to provide students with a newly-developed exam specifically designed to operationalise the CEFR levels. Hence, we resorted to combining three different assessment perspectives, i.e., self-assessment, programme tutor assessment, and assessment by seven external raters. Students and tutors employed similar CEFR grids that were selected to analyse whether the student could achieve the writing construct in question (from their knowledge of themselves or the students); raters used the Assessment grid from the Manual to rate the same students' performances from the final exam.

For the student and tutor assessment grids, we selected the following ten CEFR scales relevant for assessing writing: *Overall Written Production, Overall Written Interaction, Type of Texts, What Can They Write, Vocabulary Range & Control, Grammatical Accuracy, Orthographic Control*, *Processing Texts, Reports and Essays* and *Note Taking*. Their relevance (face validity) to this study's context was checked with two teachers on the PYP and a member of academic staff working in one of the university medical colleges.

Irrelevant scales (e.g., *Correspondence* and *Creative writing*) were excluded as they are not related to the study's context. After designing the assessment grid and before piloting, more feedback was sought from the same teachers and from colleagues from the applied linguistic field. Based on this, further scales were either eliminated or combined, e.g., *Vocabulary range* and *Vocabulary control* were combined to reduce the burden on participants (Faez et al. 2011) and therefore increase the likelihood of their engagement in the assessment. Equally, however, there was a need to ensure that relevant writing scales were covered to gather a more complete picture of the students' writing levels.

In the assessment grids, the CEFR levels A1 to C2 (including plus levels for A2, B1, and B2) were depicted as columns 1 to 9; the 10 CEFR-based categories were described in 10 separate lines, with the respective descriptors located at their correct levels (see Appendix 2). Where the CEFR scales did not contain a descriptor for the plus level, we left a blank. This basic grid was then slightly amended for the student and tutor version.

3.4.1 Students' grid

For the student grid, the "can-do" descriptors were reformulated in "I can do" statements. Using the CEFR scales based on what learners "can do" with language (CoE 2001) may improve the reliability of the findings, as using functional language (i.e., "can do" statements) has been found to increase the accuracy of self-assessment (Ross 1998). For each descriptor, students were asked to decide whether they are confident that they can perform what is depicted in the descriptor ("Yes I can"), or whether they are "not sure" that they could perform the depicted language activity. We chose the "not sure" option to allow for doubts regarding students' abilities (Alderson 2005). When students choose "Yes I can", this, in the researchers' view (by adopting a more 'conservative' approach), indicates that students are most probably able to perform the language activity depicted in that descriptor. We decided against providing a third option (e.g., "cannot do"), as this would make the analysis more complex and difficult to interpret (Ashton 2014). Figure 1 shows how the grid works.

Students are required to read the descriptors starting with *Overall Written Production*, descriptor for level A1 (1 in the grid). If they feel they can do what the descriptor states, they tick "I can do" and move on to the second descriptor, and so on until they reach a descriptor that they feel they are not sure they are capable of doing or are unable to do. In the case in Figure 1, the student ticked not sure for the descriptor at level 4 (B1). In this case, the student then proceeds to the next row (i.e., the following CEFR-based category, here *Overall Written Interaction*) and follows the same process. The student's assessment for each category is coded as the last level at which they ticked "Yes I can", in the case above the student would score 2 (A2) for *Overall Written Production*, as there is no descriptor for level A2+.

3.4.2 Tutors' grid

The tutor grid was based on the same CEFR-based grid described above. The only difference to the student grid was, that the "can do" statements were rephrased as "The student can". Tutors used the same procedure as outlined above to assess each of their students.

3.4.3 Raters' grid

Raters used the *Writing Assessment Grid* from the CEFR manual mentioned earlier (CoE 2009: 187) to assess the aforementioned student performance. We did not adapt the Grid as we wanted to use it as an independent external criterion that should reflect the CEFR construct of writing as closely as possible. Hence, the raters used the grid in its original form, encompassing the six CEFR levels (A1 to C2, without plus levels) for the five categories *Overall, Range, Coherence, Accuracy,* and *Description*.

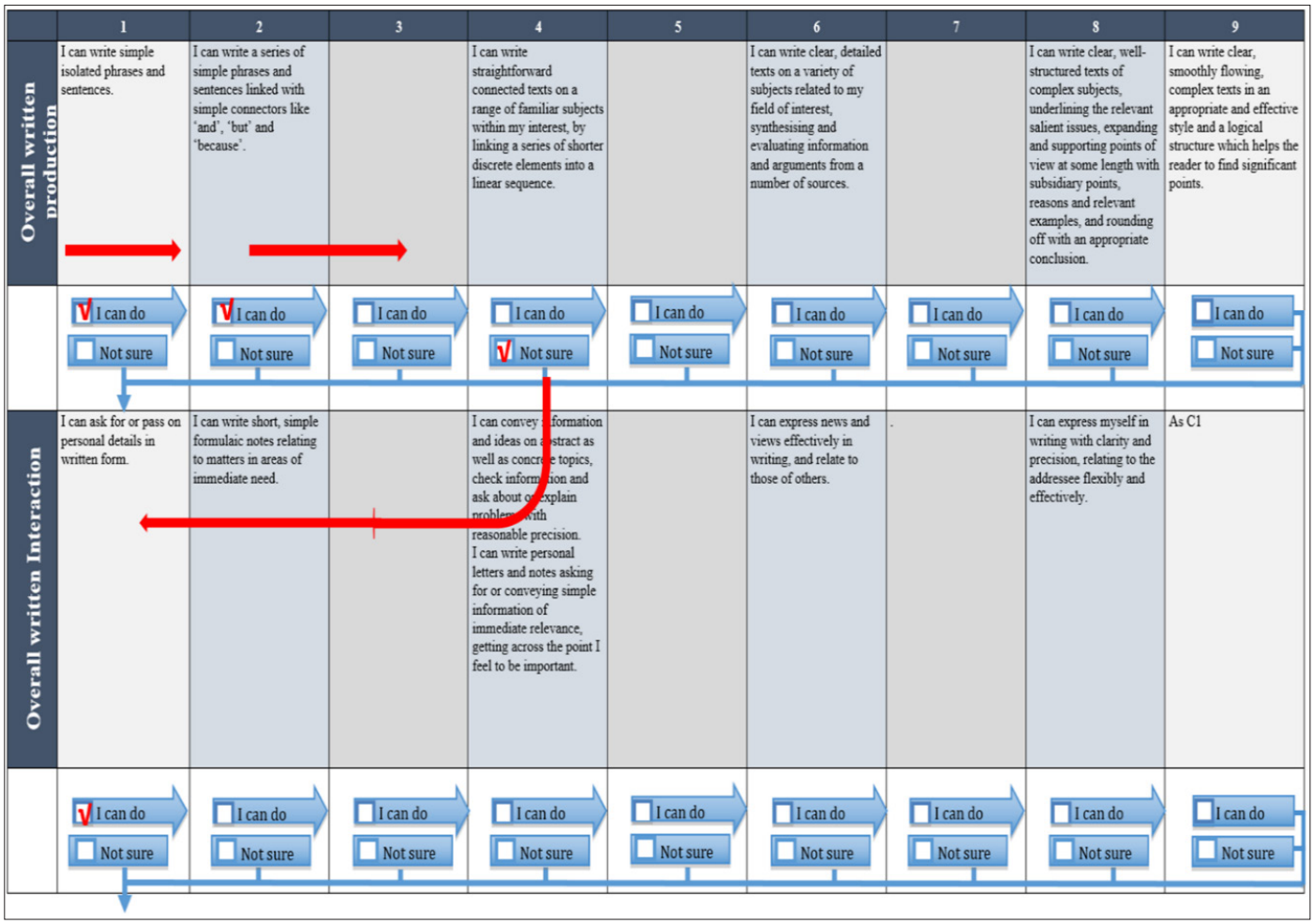


Figure 1. Student assessment grid

3.5 Data collection

Data were collected during the final stages of the PYP year, after students had taken the final PYP exam, in the expectation that participants would have developed the necessary writing skills by then.

3.5.1 Students and their tutors

All students were given a study information sheet and were familiarized with the grid. The way the CEFR scales were formatted for this study aimed to help guide students in their self-assessment, and while there was no formal training conducted to improve the reliability of assessment (Harris 1997; Little 2002; Ross 1998) nor experience in self-assessment (Engelhardt and Pfingsthorn 2013), detailed instructions were given.

Each student received her own paper-based assessment grid bearing her name and university ID (Arabic version, anonymized after data collection), so that students could be tracked, and their assessments compared with those conducted by the tutors. To mitigate against the possibility of deliberately giving inaccurate assessments of their abilities, students were encouraged to assess themselves honestly; they were reassured that their assessment would not affect any of their marks and would only be used for research purposes.

Tutors were given the same study information sheet as the students and were familiarized with the grid before using it. They received one grid for each student, containing their names and university IDs.

A total of 517 students (73 elementary, 268 intermediate and 176 advanced) submitted self-assessments and were also rated by their teachers.

3.5.2 Raters

Raters received a two-hour training session that entailed familiarisation, standardisation and benchmarking activities adapted from the manual (CoE 2009) to use the CEFR grid. After training, each of the seven raters rated the same 105 texts (the aforementioned random sample of students' performances on the PYP end-of-year exam, the same performances that had been graded by the programme tutors which yielded the ceiling effect mentioned previously). Out of these 105 students, 14 attended the elementary level of the PYP, 55 the intermediate and 36 the advanced level. The raters used the assessment grid from the manual, which originally contains the six main CEFR levels; for the data collection here, to achieve comparability with the aforementioned 9-point grid, we asked the raters to also consider the plus levels, albeit without descriptors. Raters entered their chosen levels for the five categories in a prepared excel sheet that contained these nine levels and five categories.

3.6 Methods of Analysis

We compared the results of these three perspectives (self, tutor and rater's assessments) for reliability within and between the three groups of assessors and their capability to differentiate the three PYP levels.

Cronbach's alpha showed a high reliability (of α=0.88 and α=0.95 for students and tutors' assessment, respectively), showing that the scale items measured the same underlying construct and allowed the possibility of using average scores from the ten CEFR scales (Bland and Altman 1997).

Inter-rater reliability for the five categories of the rating scale for raters was measured using Cronbach's alpha, which was also >0.8, indicating good consistency between raters, allowing to average the seven scores for each category and student.

3.6.1 RQ1

Descriptive analyses were utilized to calculate the mean and standard deviation of students' self-assessments, tutors' assessments, and raters' scores for each CEFR-based category, to ascertain whether their respective ratings yielded differences in students' performances by PYP levels.

To examine whether the differences found in the descriptive analyses are indeed significant across the three PYP levels (elementary, intermediate, and advanced), we used a one-way analysis of variance (ANOVA), as ANOVA "looks for differences between groups which are not due to chance" (Green 2013: 107). Each group of assessors was separately examined. First, the homogeneity of variance was tested (Pallant 2013). In cases where the assumption of equal variances was violated, non-parametric analysis of variance tests (i.e., the so-called Brown-Forsythe and Welch Tests, see e.g. Green 2013) were conducted. A significance level (P-value) of less than 0.05 indicates a significant difference in mean scores across the three PYP levels. In addition, the ANOVA results report η^2, which is a measure of effect size (larger effect sizes reflecting larger differences; Miles and Shevlin 2001): values around 0.02 indicate "small", values around 0.13 "medium" and values above 0.26 "large" effect sizes (Cohen 1988).

To determine the significance between each pair of the three PYP levels, we conducted post-hoc tests. If the assumption of homogeneity was met, we performed Tukey's Honestly Significant Difference (HSD) test (Pallant 2013); otherwise, for heterogeneity of variances, we used Tamhane's T2 test (Green 2013).

3.6.2 RQ2

Self-assessments and tutors' assessments were compared using a paired sample t-test (Field 2009) to identify any significant differences between the different assessments of the same students; then, correlation and agreement analyses were conducted to examine the direction and the level of agreement between these two assessor groups. To observe the strength and direction of the relationship between students' and tutors' assessments, Spearman's correlation coefficient (r) was used. Values of r of 0.00-0.19 indicate "very weak" correlation; 0.20-0.39 "weak"; 0.40-0.59 "moderate"; 0.60-0.79 "strong" and 0.80-1.0 "very strong" correlation. Additionally, the weighted Cohen's Kappa coefficient (for ordinal data such as our scores; Cohen 1968) was used to measure the degree of exact agreement between students and tutors, which takes into account the agreement that can be attributed to chance (Smeeton 1985). Kappa values of 0–0.2 indicate "slight" agreement, 0.21–0.4 "fair", 0.41–0.6 "moderate", 0.61–0.8 "substantial", 0.81–1 "almost perfect" and 1 "perfect" agreement (Landis and Koch 1977). In addition, percentages of exact agreement of student-tutor pairs were calculated, as well as agreement within one and within two adjacent CEFR levels.

For the 105 cases where three sets of data existed, we performed ANOVA, correlation and post-hoc tests, to compare the means of the self-assessments, tutors' assessments, and scores given by the external raters for the same students. This allowed for the examination of the direction and relation among the assessments provided by these three groups.

4 Results

4.1 RQ1 CEFR writing levels assessed by students, tutors, and raters separately across the three PYP levels

4.1.1. Descriptive Analysis

First, we present the results of the descriptive analyses (mean and standard deviation (SD)) for the three PYP levels (elementary, intermediate, and advanced), as perceived by students' self-assessment, tutors' and raters' assessments. Table 1 illustrates the self-assessment results, the results for tutors and raters are presented in Appendix 3 for space reasons.

Table 1. ***Descriptive analysis of PYP students' self-assessment across the PYP levels***

	Elementary n=73		Intermediate n=268		Advanced n=176	
CEFR Categories	M	SD	M	SD	M	SD
Overall Written Production	5.57	2.35	6.24	2.17	7.91	1.66
Overall Written Interaction	3.93	2.10	4.22	2.28	6.67	2.56
Type of Texts	3.94	2.05	4.28	2.23	6.27	2.48
What Can They Write	4.40	2.24	4.87	2.25	6.80	1.97
Vocabulary Range & Control	3.55	2.00	3.95	1.97	5.85	2.37
Grammatical Accuracy	4.32	2.68	5.08	2.39	6.16	2.84
Orthographic Control	5.05	2.77	5.41	2.67	7.00	2.14
Processing Texts	3.81	1.54	4.39	1.76	6.13	2.23
Reports and Essays	4.14	2.44	4.50	2.41	6.75	2.04
Note Taking	5.22	2.48	5.44	2.30	6.94	2.17
Average of Scales	4.48	1.58	4.92	1.53	6.73	1.43
M=Mean, SD=Standard deviation Coding scheme for CEFR categories: 1 (A1); 2 (A2); 3 (A2+); 4 (B1), 5 (B1+); 6 (B2); 7 (B2+); 8 (C1); 9 (C2)						

For each category and each group of assessors, mean scores increased from elementary to intermediate to advanced level students, indicating that the three group of assessors could differentiate between the three PYP levels, unlike the end-of-year exam.

4.1.2. ANOVA

To find out whether the increase across the three PYP levels is significant, we conducted ANOVA analyses. While we present the results for the three assessor groups here, the supporting tables are presented in the appendix for space reasons: Appendix 4 contains the tables for students; Appendix 5 for tutors and Appendix 6 for raters.

Looking at the students' self-assessment across the three PYP levels (Appendix 4, Tables 6 [ANOVA] and 7 [non-parametric analysis of variance tests]), the effect sizes were 0.095 to 0.26, indicating medium-to-large effect sizes for the differences between elementary, intermediate and advanced groups. The largest effect size was observed for the average of all categories (η^2=0.26). From the post hoc pairwise results (Appendix 4, Tables 8 [Tukey] and 9 [Tamhane]), significant differences were evident between the advanced and intermediate levels and the advanced and elementary levels. There were no significant differences between the elementary and intermediate levels, except in the *Processing Texts* category, where the scores for students from all three levels differed significantly from each other.

With regard to tutors' assessment, there were significant differences for all CEFR categories across the three PYP levels (Appendix 5, Tables 10 and 11). A substantial effect (η^2) was observed in most categories, except for *Note Taking*, where the effect was comparatively small. The results of the post-hoc tests (Appendix 5, Tables 12 and 13) showed significant differences in tutors' assessments between all three PYP levels, in the expected directions, with the elementary level receiving significantly lower scores compared to the intermediate level, and the intermediate level significantly lower than the advanced level.

When it comes to the external raters, we used the average scores across the seven raters (Appendix 6). The ANOVA (Table 14) showed significant differences across the three PYP levels, with large effect sizes in the expected directions (i.e., the elementary level receiving significantly lower scores compared to the intermediate level, and the intermediate level scoring significantly lower than the advanced level). The post-hoc analysis (Appendix 6, Table 15) showed significant differences in the raters' scores of students at the advanced versus intermediate or elementary levels for all categories (*Range, Coherence, Accuracy, Description* and *Overall*), but not between the intermediate and elementary levels in any category.

4.2 RQ2 comparing the three participating assessor groups: students, tutors and raters

RQ2 examined the extent to which the three participating assessor groups (students, tutors, raters) are comparable in their assessment using the selected CEFR-based categories. As two groups (students and tutors) used the same tool for assessment, we first compared these two groups, using a paired sample t-test to check whether the PYP students' and tutors' assessments differed significantly. Then, a comparison across the three groups was conducted, using correlations and ANOVA to compare the means between self-, tutors' and raters' scores of the same 105 students.

4.2.1. Self- and tutors' assessments

We compared means for students and tutors using the paired t-test. Cohen's d provides an estimate of the effect size (Pallant 2013), where d=0.2 is considered "small", 0.5 "medium" and 0.8 "large" (Cohen 1988). Appendix 7, Table 16 contains the detailed results.

At the elementary level, the largest effect sizes were observed for *Overall written production* and *Processing texts*, followed by *Note taking*, with students rating themselves significantly higher than their tutors. At the intermediate level, the largest (medium size) differences were for *Type of texts, Overall written interaction*, and *Vocabulary range and control*; in each case the students rated themselves lower than the tutors.

With the advanced-level students, scores on most of the CEFR-based categories showed very similar means (with non-significant P-values and small effect sizes), indicating that students and their tutors have similar perceptions of the CEFR levels students have reached in those categories. However, this was not true for all scales, with tutors scoring significantly higher for *Type of texts* and significantly lower for *Note taking* and *Reports and essays* (small effect size).

Appendix 7, Table 17 shows the correlation between students and teachers' scores, the weighted kappa (measure of agreement) and the percentages of scores with exact agreement (identical level assigned), or agreements within one or two levels.

There was a significant positive correlation between the scores of students and their tutors for all CEFR-based categories, though the strength of the relation was weak to moderate (all r<0.30 for individual items; r=0.39 for overall average). Weighted Kappa was low (max=0.39), indicating only weak to moderate agreement in students' and tutors' assessment. Overall, 19.0% of pairs agreed exactly; 52.4% agreed within one level and 79.9% within two levels, showing fairly close agreement between the tutors' assessment and their students' self-assessment.

4.2.2. Self-, tutors' and raters' assessments

Students', tutors' and raters' assessments were compared using only the sample where data exist from self-assessment, tutor assessment and mean scores across the seven raters (n=105, including all three levels). Correlation analysis was carried out to explore the relations between the three assessments (students, tutors and raters). Table 2 presents the results.

Table 2. ***Overall correlation analysis between self, tutors' and raters' assessment***

	Raters n=105		Students n=105		Tutors n=105	
	Pearson Correlation	P-value	Pearson Correlation	P-value	Pearson Correlation	P-value
Raters	1		0.44**	<0.001	-0.11	0.27
Students	0.44**	<0.001	1		-0.065	0.51
Tutors	-0.11	0.27	-0.065	0.51	1	

** Correlation is significant at the .01 level (2-tailed).

There exists a significant positive correlation between the raters' scores and the students' self-assessment, although the tutor's scores did not correlate significantly with either students' or raters' scores. The patterns of averaged scores from self-assessment for elementary, intermediate and advanced levels were B1, B1 and B2 (i.e., elementary and intermediate scored the same, then up one level for advanced students), and for raters the pattern was similar: A2+, A2+ and B1 (i.e., elementary and intermediate scored the same, then up one level for advanced students). However, the pattern for teachers' ratings differed: A2+, B1 and B2, respectively.

To compare the three groups, a one-way ANOVA was used. Table 3 shows the ANOVA results, comparing students', tutors' and raters' assessments for the 105 participants for whom all three types of assessments were available.

Table 3. ***One-way ANOVA between students, tutors and raters***

	SS	df	MS	F	P-value	η^2
Between Groups	113.74	2	56.87	25.99	p<.0001	0.20
Within Groups	662.9	303	2.19			
Total	776.65	305				

SS=Sum of squares, df=degrees of freedom, MS=mean square, F=F ratio, η2=Effect size:0.02=small; 0.13=medium; 0.26=large.

We found significant differences in the scores between self-assessment, tutors' and raters' assessments, with a large effect size. To identify where the differences were located, Tukey's post hoc analysis was conducted (Appendix 8, Table 18), which showed that the raters gave significantly lower scores than both the students and tutors, and this was true across all three PYP levels (Appendix 8, Table 19).

5 Discussion and conclusion

This study aimed to explore assessments by three groups of assessors, i.e., students, their tutors and external raters, in order to yield assessment approaches that would be able to differentiate between the three proficiency levels taught at the PYP (Intensive English) programme for medical students. At the same time, we sought to benchmark the three PYP proficiency levels achieved in writing at the end of the PYP to a recognized framework (the CEFR). We also aimed to deepen our understanding of self-assessment, tutor assessment, and scores of independent raters based on relevant CEFR scales in the Saudi Arabian higher education context.

5.1 Research Question 1

Our first research question, i.e., "Can students' self-assessment, tutors' assessment, and raters' assessment (using selected CEFR scales) reliably differentiate students' writing proficiency among the three PYP levels?" was partially supported. The students placed in elementary level generally received lower scores compared to those at the intermediate level, and the intermediate level students scored lower than the advanced level students; differences were significant between advanced and intermediate students, and between advanced and elementary students, although the differences between elementary and intermediate students were less pronounced.

The CEFR can potentially be used to gain a criterion-referenced general overview of the students' proficiency levels as a starting point in a context outside of Europe such as Saudi Arabia, with participants having no or little experience with using the CEFR scales (Abdulhaleem and Harsch 2018). The scores could be benchmarked against a recognised framework (i.e., the CEFR), although only selected scales of the CEFR were used in the assessment grids. Scores for elementary, intermediate and advanced level students' self-assessments were equivalent to CEFR levels B1, B1 and B2; scores from tutor assessment placed students at A2+, B1+ and B2 respectively, while the external raters placed students at A2+, A2+ and B1. We will discuss the meaning of these results below, when taking a closer look at agreement levels.

5.2 Research Question 2

Our second research question, i.e., "To what extent are the scores from the three assessor groups comparable, taking the three PYP levels into account?" was also partially supported.

When comparing **students and tutors**, a moderate yet significant correlation between the students' self-assessments and tutors' assessments was found (r=0.39). This is similar to the average correlation identified by Falchikov and Boud (1989), in their meta-analysis of studies comparing self-assessment with teachers' marks, which also reported an average correlation of r=0.39.

Even if results correlate significantly, this does not necessarily demonstrate exact or close agreement (Fleiss and Cohen 1973; Cohen 1968). To the best of our knowledge, few studies investigating self-assessment – especially language proficiency-related studies – have compared agreement between students' self-assessment and their tutors' assessment. In this study, we used a weighted kappa to test the significance and percentage agreement between the two assessments. Exact agreement between students' and tutors' assessment was low (19%) but was higher between one (52.4%) and two (79.9%) adjacent CEFR scores. The two adjacent scores in the study means that the agreement is equal to "one and a half levels, e.g., A2+ to B1+", which is considered sufficient agreement according to the CEFR manual (CoE 2009: 37). This means that the students were not too far away in their perceptions of their CEFR levels from those of their tutors, suggesting the value of using the CEFR scales as exemplified in this study.

Looking at the three PYP proficiency levels separately, elementary students self-assessed their CEFR levels as B1, tutors assessed them as A2+. So elementary-level students tend to overestimate their proficiency (CEFR levels) compared to tutors. This was expected, as it has been widely found in the literature that low-proficiency students tend to overestimate their proficiency (Babaii et al. 2016; Leach 2012; Ünaldı 2016; Blue 1988).

Intermediate students achieved levels of B1 by self-assessment and B1+ by tutors. In contrast to the elementary level students, some intermediate-level students were found to underestimate their proficiency compared to their tutors' assessment. Similar results were also found in the literature, where higher proficiency students show a tendency to underestimate their proficiency level when they assess themselves (Kruger and Dunning 1999; Hodges et al. 2001; Lejk and Wyvill 2001; Tejeiro et al. 2012).

Advanced-level students achieved B2 according to self- and tutor-assessment. Generally, their self-assessment was closer to that of their tutors and showed less variance than at the other levels, indicating more accurate self-assessment. This was found in other studies that described more similarities between the students and their teachers' marks/assessment and therefore considered the assessment as more accurate when students came from higher-proficient levels (Kun 2016; Sahragard and Mallahi 2014; Ünaldı 2016), possibly due to the Dunning-Kruger effect, where students at higher proficiency levels have the cognitive ability to assess and judge their proficiency more accurately**.**

With regard to comparing **students and raters**, there was a significant moderate correlation between the students' self-assessments' and raters' assessments (r=0.44). The pattern of levels assigned by students at each of the proficiency levels (B1 and B2 for elementary, intermediate and advanced) was similar to that assigned by the raters (A2+, A2+ and B1, respectively), although the raters' assessments were around one CEFR level lower than the students' assessments across all PYP proficiency levels. These findings are consistent with those of Fleckenstein et al. (2018).

Comparing **tutors and raters**, agreement between these two groups was lower than between students and teachers or students and raters. Different explanations can be given for the discrepancies between the tutors' assessment and the raters' scores. One explanation is that though the tutors are following criterion-referenced assessment as it is usually the case when using the CEFR scales (Fleckenstein et al. 2018; Hughes 2002), there is still the possibility that the tutors tended to compare the students within or between their classes (norm-referenced assessment) (Fleckenstein et al. 2018; Lok et al. 2016). However, the grades assigned by the tutors were the most discriminating (different average CEFR levels assigned to elementary, intermediate and advanced level students), whereas students and raters gave the same levels to elementary and intermediate students.

Moreover, the raters were focusing on a small sample of specific exam texts, which may be easier to judge than students' proficiency in general (as for students and tutors using the CEFR scales) (Fleckenstein et al. 2018; Südkamp et al. 2012), However, raters only scored the end-of-year exam texts, which could have been inadequate to demonstrate students' full range of writing proficiency, as for example, level C1 requires complex subjects, a wide range of topics and imaginative texts, whereas the exam (based at A2 level) only required students to write 120 words in 60 minutes on a general topic

about their daily routine at the university, with little scope to demonstrate higher skills. There may be a difference between what students and their teachers assess they "can do" in general and what they actually were able to demonstrate in the exam. Another source of variance is to be found in the grid the raters used, which may have been inappropriate for the exam at hand, or the rater training may have been inadequate.

5.3 Conclusions

Based on our findings, and despite the limitations identified, there are indications enough to argue for the usefulness of the CEFR to identify students' proficiency levels. Students and tutors could potentially use the CEFR-based grids and compare their respective assessments as a basis for identifying areas on which to focus for further learning. Considering that the participating students and tutors had not been extensively trained in using the CEFR scales to identify students' proficiency levels in writing, the findings for correlations and underestimation and overestimation of self-assessment are similar to those found in the literature. As mentioned in Moonen et al. (2013), many people have little experience of and exposure to the use of the CEFR scales, and as suggested by Davis (2015), Fahim and Bijani (2011), Fleckenstein et al. (2018), and Weigle (1994), with proper instruction and training, the tutors and students might be more accurate in their assessment.

The study findings revealed noticeable variations in the average scores across the three PYP levels in the assessments conducted by students, tutors, and raters. These disparities provide insights into the applicability of the CEFR scales. Furthermore, the results highlight that the CEFR can serve as a valuable criterion-referenced tool for gaining a broad understanding of students' writing proficiency levels, even within a non-European setting where participants may possess limited familiarity with the CEFR scales. This serves as a foundation for future assessment and evaluation endeavors, encouraging further exploration in this area.

6 References

Abdulhaleem, Ebtesam & Claudia Harsch. 2018. Using the CEFR scales to assess students' proficiency levels in a Saudi-Arabian higher education context. In Anikó Brandt, Astrid Buschmann-Göbels & Claudia Harsch. (eds.). *Der Gemeinsame Europäische Referenzrahmen für Sprachen und seine Adaption im Hochschulkontext. Erträge des 6. Bremer Symposions*. Fremdsprachen in Lehre und Forschung Bd. 51. [*The Common European Framework of Reference for Languages and its adaptation for higher education contexts. Proceedings of the 6th Bremer Symposium*. Foreign Languages in Teaching and Research 51.] Bochum: AKS. 167-178.

Alderson, J. Charles. 2005. *Diagnosing foreign language proficiency: The interface between learning and assessment.* London: Continuum.

Aljumah, Fahad Hamad. 2012. Saudi learner perceptions and attitudes towards the use of blogs in teaching English writing course for EFL majors at Qassim University. *English Language Teaching* 5(1). 100-116. https://doi.org/10.5539/elt.v5n1p100.

Alkubaidi, Miriam A. 2014. The relationship between Saudi English major university students' writing performance and their learning style and strategy use. *English Language Teaching* 7(4). 83-95. https://doi.org/10.5539/elt.v7n4p83.

Ashton, Karen. 2014. Using self-assessment to compare learners' reading proficiency in a multilingual assessment framework. *System*, 42, 105–119. https://doi.org/10.1016/j.system.2013.11.006.

Atai, Mahmood Reza & Leila Shoja. 2011. A triangulated study of academic language needs of Iranian students of computer engineering: Are the courses on track? *RELC Journal* 42(3). 305-323. https://doi.org/10.1177/0033688211419392.

Babaii, Esmat, Shahin Taghaddomi & Roya Pashmforoosh. 2016. Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing* 33(3). 411-437. https://doi.org/10.1177/0265532215590847.

Bérešová, Jana. 2011. The impact of the common European framework of reference on teaching and testing in Central and Eastern European context. *Synergies Europe* 6. 177-190. https://gerflint.fr/Base/Europe6/jana.pdf (accessed 22 January 2025).

Bland, J. Martin & Douglas G. Altman. 1997. Statistics notes: Cronbach's alpha. *The British Medical Journal*, 314, 572. https://doi.org/10.1136/bmj.314.7080.572.

Blue, George M. 1988. Self-assessment: The limits of learner independence. Individualization and autonomy in language learning. *ELT Documents* 131. 100-118.

Boud, David. 1991. *Implementing student self-assessment.* Campbelltown, N. S. W.: Higher Education Research and Development Society of Australasia (HERDSA).

Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4). 213–220. https://doi.org/10.1037/h0026256. PMID 19673146

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Council of Europe. 2001. *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 (CEFR).* Strasbourg: Council of Europe.

Davis, Larry. 2015. The influence of training and experience on rater performance in scoring spoken language. *Language Testing* 33(1). 117-135. https://doi.org/10.1177/0265532215582282.

Dragemark-Oscarson, Anne 2009. *Self-assessment of writing in learning English as a foreign language. A study at the upper secondary school level.* (Doctoral thesis, University of Gothenburg). http://hdl.handle.net/2077/19783 (accessed 22 January 2025).

Dunning, David, Chip Heath & Jerry M. Suls. 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest* 5(3). 69-106. https://doi.org/10.1111/j.1529-1006.2004.00018.x.

Ebad, Ryhan. 2014. The role and impact of English as a language and a medium of instruction in Saudi higher education institutions: Students-instructors perspective. *Studies in English Language Teaching* 2(2). 140-148. https://doi.org/10.22158/selt.v2n2p140.

Engelhardt, Maike & Joanna Pfingsthorn. 2013. Self-assessment and placement tests – a worthwhile combination? *Language Learning in Higher Education* 2(1). 75-89. https://doi.org/10.1515/cercles-2012-0005.

Faez, Farahnaz, Shelley Taylor, Suzanne Majhanovich, Patrick Brown & Maureen Smith. 2011. Teachers' reactions to CEFR's task-based approach for FSL classrooms. *Synergies Europe* 6. 109-120. https://gerflint.fr/Base/Europe6/faez.pdf (accessed 22 January 2025).

Fahim, Mansoor & Houman Bijani. 2011. The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing* 1(1). 1-16.

Falchikov, Nancy & David Boud. 1989. Student self-assessment in higher education: A meta-analysis. *Review of Educational Research* 59. 395–430. https://doi.org/10.3102/00346543059004395.

Field, Andy. 2009. *Discovering statistics using SPSS* (3rd ed.). Los Angeles: Sage Publications.

Fleckenstein, Johanna, Michael Leucht & Olaf Köller. 2018. Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Language Assessment Quarterly* 15(1). 90-101. https://doi.org/10.1080/15434303.2017.1421956.

Fleiss, Joseph L. & Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33(3). 613-619. https://doi.org/10.1177/001316447303300309.

Green, Anthony. 2012. *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range* (Vol. 2). Cambridge: Cambridge University Press.

Green, Rita. 2013. *Statistical analyses for language testers.* Basingstoke, UK: Palgrave Macmillan.

Harris, Michael. 1997. Self-assessment of language learning in formal settings. *ELT Journal* 51(1). 12–20. https://doi.org/10.1093/elt/51.1.12.

Harsch, Claudia. 2018. How suitable is the CEFR for setting university entrance standards? *Language Assessment Quarterly* 15(1). 102-108. https://doi.org/10.1080/15434303.2017.1420793.

Hellmann, Kate. 2013. *What do I need to succeed: The case of Arab engineering graduate students' self-perceptions of writing* (Doctoral dissertation, University of Idaho).

Hodges, Brian, Glenn Regehr & Dawn Martin. 2001. Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine* 76(10). S87-S89. https://doi.org/10.1097/00001888-200110001-00029.

Hughes, Arthur. 2002. *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Kruger, Justin & David Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77(6). 1121-1134. https://doi.org/10.1037//0022-3514.77.6.1121.

Kun, András István. 2016. A comparison of self versus tutor assessment among Hungarian undergraduate business students. *Assessment & Evaluation in Higher Education* 41(3). 350-367. https://doi.org/10.1080/02602938.2015.1011602.

Leach, Linda. 2012. Optional self-assessment: Some tensions and dilemmas. *Assessment & Evaluation in Higher Education* 37(2). 137-147. https://doi.org/10.1080/02602938.2010.515013.

Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1). 159-174. https://doi.org/10.2307/2529310.

Lejk, Mark & Michael Wyvill. 2001. The effect of the inclusion of self-assessment with peer assessment of contributions to a group project: A quantitative study of secret and agreed assessments. *Assessment & Evaluation in Higher Education* 26(6). 551-561. https://doi.org/10.1080/02602930120093887.

Levene, H. 1960. Robust tests for equality of variance. In Ingram Olkin (ed.), *Contributions to Probability and Statistics,* 278–292. Palo Alto, CA: Stanford University Press.

Little, David. 2002. The European language portfolio: Structure, origins, implementation and challenges. *Language Teaching* 35(3). 182-189. https://doi.org/10.1017/S0261444802001805.

Little, David. 2007. The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *Modern Language Journal* 91(4). 645-655. https://doi.org/10.1111/j.1540-4781.2007.00627.x.

Lok, Beatrice, Carmel McNaught & Kenneth Young. 2016. Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. *Assessment & Evaluation in Higher Education* 41(3). 450-465. https://doi.org/10.1080/02602938.2015.1022136.

McMullen, Maram George. 2009. Using language learning strategies to improve the writing skills of Saudi EFL students: Will it really work? *System: An International Journal of Educational Technology and Applied Linguistics* 37(3). 418-433. https://doi.org/10.1016/j.system.2009.05.001.

McMullen, Maram George. 2014. The value and attributes of an effective preparatory English program: Perceptions of Saudi university students. *English Language Teaching* 7(7). 131-140. https://doi.org/10.5539/elt.v7n7p131.

Miles, Jeremy & Mark Shevlin. 2001. *Applying regression and correlation: A guide for students and researchers.* London: Sage Publications Ltd.

Moonen, Machteld, Evelien Stoutjesdijk, Rick de Graaff & Alexxandra Corda. 2013. Implementing the CEFR in secondary education: Impact on FL teachers' educational and assessment practice. *International Journal of Applied Linguistics* 23(2). 226–246. https://doi.org/10.1111/ijal.12000.

North, Brian. 2014. *The CEFR in practice.* Cambridge: Cambridge University Press.

Obeid, Rana. 2017. Second language writing and assessment: Voices from within the Saudi EFL context. *English Language Teaching* 10(6). 174-181. https://doi.org/10.5539/elt.v10n6p174.

Oraif, Iman. 2016. The right approach in practice: A discussion of the applicability of EFL writing practices in a Saudi context. *English Language Teaching* 9(7). 97-102. https://doi.org/10.5539/elt.v9n7p97.

Oxford University Press. 2001. *Quick placement test: Paper and pen test, user manual*.

Pallant, Julie. 2013. *SPSS survival manual* (5th ed.). Maidenhead, UK: McGraw-Hill Education.

Paris, Scott G. & Peter Winograd. 1990. How metacognition can promote academic learning and instruction. In Beau Fly Jones & Lorna Idol. (eds.). *Dimensions of Thinking and Cognitive Instruction,* 15-51. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Punch, Keith F. 2005. *Introduction to Social Research: Quantitative and Qualitative Approaches* (2nd ed.). London: Sage Publications Ltd.

Ross, Steven. 1998. Self-assessment in Second Language Testing: A Meta-analysis and Analysis of Experiential Factors. *Language Testing* 15(1). 1-20. https://doi.org/10.1177/026553229801500101.

Sahragard, Rahman & Mallahi, Omid. 2014. Relationship between Iranian EFL learners' language learning styles, writing proficiency and self-assessment. *Procedia-Social and Behavioral Sciences* 98. 1611-1620.

Shukri, Nadia Ahmad. 2014. Second language writing and culture: Issues and challenges from the Saudi learners' perspective. Arab World English Journal, 5(3), 190–207. https://awej.org/second-language-writing-and-culture-issues-and-challenges-from-the-saudi-arners-perspective (accessed 22 January 2025).

Smeeton, Nigel C. 1985. Early history of the kappa statistic. *Biometrics* 41(3). 795.

Südkamp, Anna, Johanna Kaiser & Jens Möller. 2012. Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology* 104(3). 743-762. https://doi.org/10.1037/a0027627.

Tannenbaum, Richard J., & E. Caroline Wylie. 2005. Mapping English language proficiency test scores onto the common European framework. *ETS Research Report Series* RR-05-18, TOEFL-RR-80. Princeton, NJ: Educational Testing Service.

Tejeiro, Ricardo A., Jorge L. Gómez-Vallecillo, Antonio F. Romero, Manuel Pelegrina, Agustín Wallace & Enrique Emberley. 2012. Summative self-assessment in higher education: Implications of its counting towards the final mark. *Electronic Journal of Research in Educational Psychology* 10(2). 789-812.

Ünaldı, İhsan. 2016. Self and teacher assessment as predictors of proficiency levels of Turkish EFL learners. *Assessment & Evaluation in Higher Education* 41(1). 67-80. https://doi.org/10.1080/02602938.2014.980223.

Weigle, Sara Cushing. 1994. Effects of training on raters of ESL compositions. *Language Testing* 11(2). 197-223. https://doi.org/10.1177/026553229401100206.

7 Biography

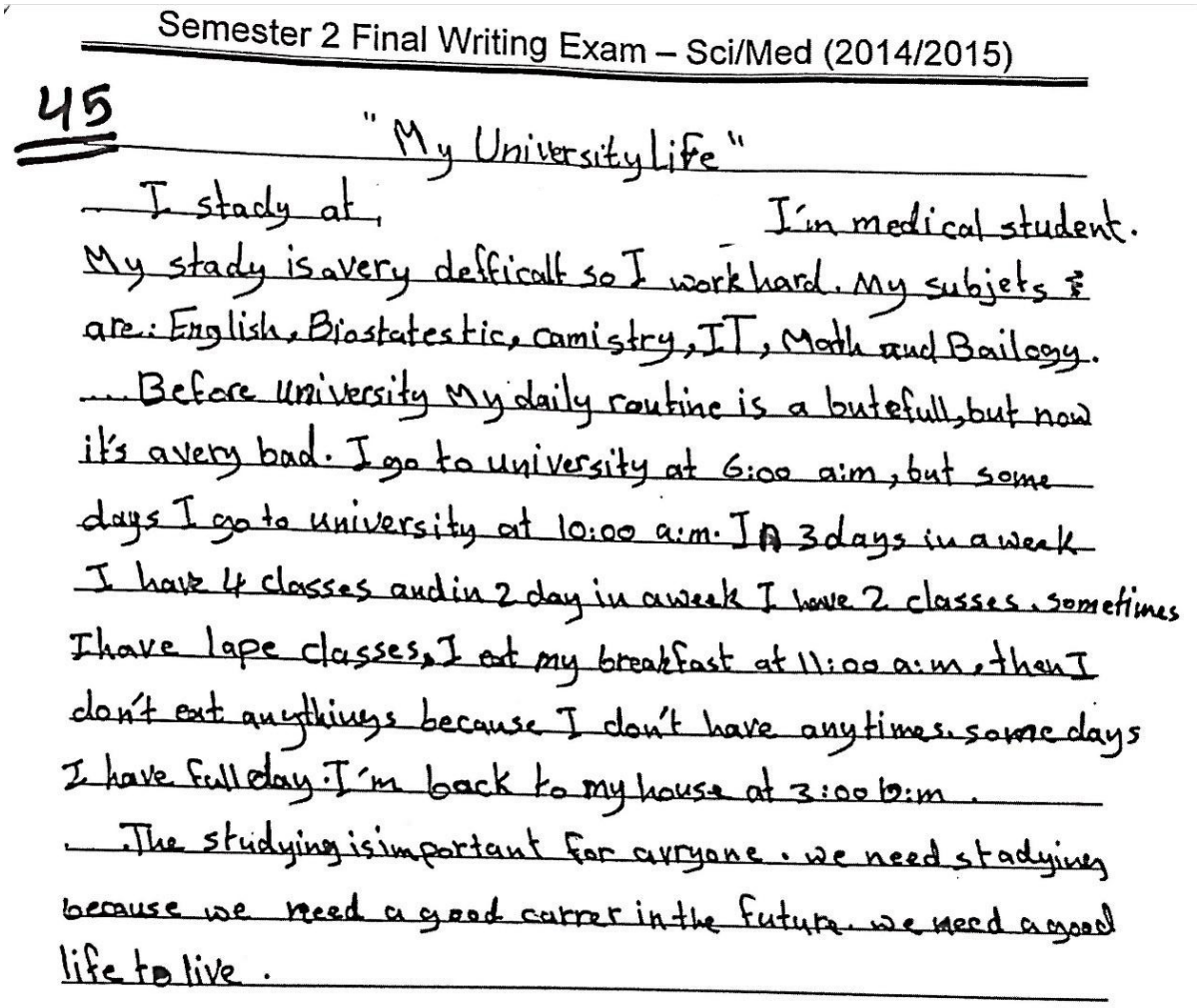
Ebtesam Abdulhaleem is an Assistant Professor and a former staff member at King Saud University. Currently, she is the Head of the Testing Development Unit at King Salman Global Academy for Arabic Language. Abdulhaleem holds a PhD in Applied Linguistics from the University of Warwick and a postgraduate certificate in language testing from Lancaster University. Her research interests encompass various areas, including language testing and assessment, assessment literacy, corpus linguistics, the role of practitioners as researchers, as well as well-being and professional development.

Claudia Harsch is a professor at the University of Bremen, specialising in language learning, teaching and assessment. She has worked in Germany and in the UK, and is active in teacher training worldwide.

Her research interests focus on areas such as language assessment, language and migration, the development of language assessment literacy, and the implementation of the CEFR. Claudia is currently the immediate past president of the International Language Testing Association (president from 2023-24), and was president of the European Association of Language Testing and Assessment from 2016-2019.

8 Appendices

8.1 Appendix 1: Samples of students' written texts in the end of year exam



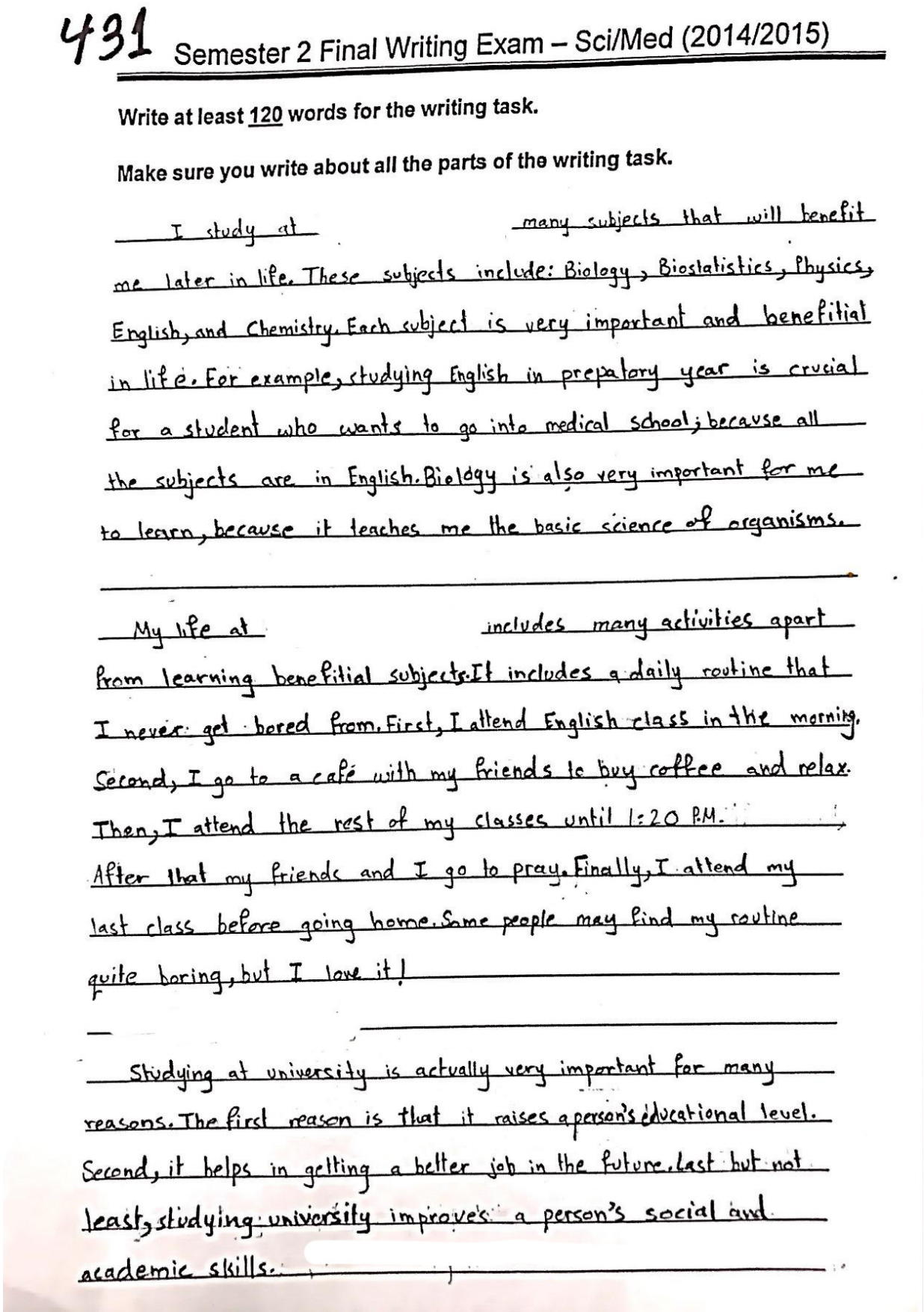
Semester 2 Final Writing Exam – Sci/Med (2014/2015)

45

"My University Life"

I stady at, I'm medical student. My stady is avery defficalt so I work hard. My subjets are: English, Biostatestic, camistry, IT, Math and Bailogy. Before university my daily routine is a butefull, but now it's avery bad. I go to university at 6:00 a:m, but some days I go to university at 10:00 a:m. In 3 days in a week I hawe 4 classes and in 2 day in aweek I howe 2 classes. sometimes Ihave lape classes, I eat my breakfast at 11:00 a:m, than I don't eat anythings because I don't have anytimes. some days I have full day. I'm back to my house at 3:00 p:m. The studying is important for avryone. we need stadying because we need a good carrer in the futura. we need a good life to live.

Figure 2a: Sample one of students' written texts from the end-of-year final exam



431 Semester 2 Final Writing Exam – Sci/Med (2014/2015)

Write at least 120 words for the writing task.

Make sure you write about all the parts of the writing task.

I study at many subjects that will benefit me later in life. These subjects include: Biology, Biostatistics, Physics, English, and Chemistry. Each subject is very important and benefitial in life. For example, studying English in prepatory year is crucial for a student who wants to go into medical school; because all the subjects are in English. Biology is also very important for me to learn, because it teaches me the basic science of organisms.

My life at includes many activities apart from learning benefitial subjects. It includes a daily routine that I never get bored from. First, I attend English class in the morning. Second, I go to a café with my friends to buy coffee and relax. Then, I attend the rest of my classes until 1:20 P.M. After that my friends and I go to pray. Finally, I attend my last class before going home. Some people may find my routine quite boring, but I love it!

Studying at university is actually very important for many reasons. The first reason is that it raises a person's educational level. Second, it helps in getting a better job in the future. Last but not least, studying university improves a person's social and academic skills.

Figure 2b: Sample two of students' written texts from the end-of-year final exam

8.2 Appendix 2. The student assessment grid

Reports and essays	No descriptors available. Start from number 4	No descriptors available. Start from number 4	No descriptors available. Start from number 4	I can write very brief reports to a standard conventionalized format, which pass on routine factual information and state reasons for actions.	I can write short, simple essays on topics of interest. I can summarize report and give his/her opinion about accumulated factual information on familiar routine and non-routine matters within my field with some confidence.	I can write an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. I can synthesise information and arguments from a number of sources.	I can write an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. I can evaluate different ideas or solutions to a problem.	I can write clear, well-structured expositions of complex subjects, underlining the relevant salient issues. I can expand and support points of view at some length with subsidiary points, reasons and relevant examples.	I can produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literacy works. I can provide an appropriate and effective logical structure which helps the reader to find significant points.
	I can do / Not sure	I can do / Not sure	Can do / Not sure	Can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure
Note Taking	No descriptors available. Please start reading from column number 4.	No descriptors available. Please start reading from column number 4.	No descriptors available. Please start reading from column number 4.	I can take notes as a list of key points during a straightforward lecture, provided the topic is familiar, and the talk is both formulated in simple language and delivered in clearly articulated standard speech.	I can take notes during a lecture which are precise enough for my own use at a later date, provided the topic is within my field of interest and the talk is clear and well-structured.	I can understand a clearly structured lecture on a familiar subject, and can take notes on points which strike me as important, even though I tend to concentrate on the words themselves and therefore to miss some information.		I can take detailed notes during a lecture on topics in my field of interest, recording the information so accurately and so close to the original that the notes could also be useful to other people.	I am aware of the implications and allusions of what is said and can make notes on them as well as on the actual words used by the speaker.
	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure
Orthographic control	I can copy familiar words and short phrases e.g. simple signs or instructions, names of everyday objects, names of shops and set phrases used regularly. I can spell my address, nationality and other personal details.	I can copy short sentences on everyday subjects – e.g. directions how to get somewhere. I can write with reasonable phonetic accuracy (but not necessarily fully standard spelling) short words that are in my oral vocabulary.		I can produce continuous writing which is generally intelligible throughout. Spelling, punctuation and layout are accurate enough to be followed most of the time.		I can produce clearly intelligible continuous writing which follows standard layout and paragraphing conventions. My spelling and punctuation are reasonably accurate but may show signs of mother tongue influence.		My layout, paragraphing and punctuation are consistent and helpful. My spelling is accurate, apart from occasional slips of the pen.	My writing is orthographically free of error.
	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure
Processing texts	I can copy out single words and short texts presented in standard printed format.	I can copy out short texts in printed or clearly handwritten format.	I can pick out and reproduce key words and phrases or short sentences from a short text within the learner's limited competence and experience.	I can collate short pieces of information from several sources and summarize them for somebody else. I can paraphrase short written passages in a simple fashion, using the original text wording and ordering.		I can summarize a wide range of factual and imaginative texts, commenting on and discussing contrasting points of view and the main themes. I can summarize extracts from news items, interviews or documentaries containing opinions, argument and discussion. I can summarize the plot and sequence of events in a film or play.		I can summarize long, demanding texts.	I can summarize information from different sources, reconstructing arguments and accounts in a coherent presentation of the overall result.
	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure	I can do / Not sure

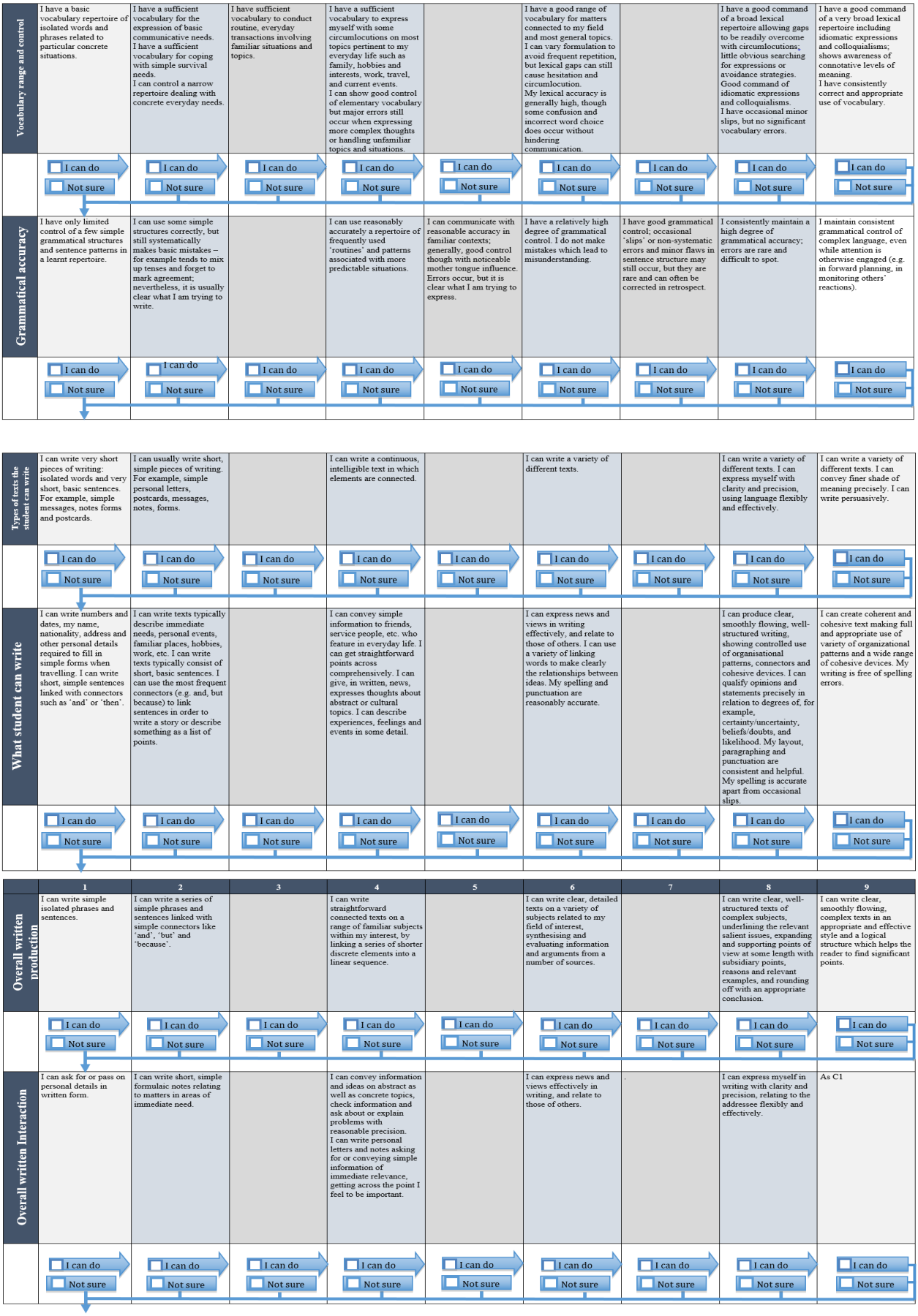


Figure 3 shows the 10 scales of the assessment grid that students were asked to complete.

8.3 Appendix 3.

Descriptive statistics for self-assessment, tutor assessment and rater scores

Tables 4 and 5 show the means and standard deviations for the scores for teacher- and rater-assessments, respectively.

Table 4. ***Descriptive analysis of PYP tutors' assessment across PYP levels***

	Elementary n=73		Intermediate n=268		Advanced n=176	
CEFR Categories	M	SD	M	SD	M	SD
Overall Written Production	4.38	1.86	5.99	2.05	7.56	1.89
Overall Written Interaction	4.12	2.03	5.63	2.16	6.88	1.82
Type of Texts	4.46	2.24	5.80	2.21	7.28	1.80
What Can They Write	3.52	1.85	4.98	1.85	6.52	2.04
Vocabulary Range & Control	3.80	1.59	4.96	1.82	6.31	2.19
Grammatical Accuracy	3.88	1.89	4.98	1.74	6.16	2.24
Orthographic Control	4.22	2.52	4.89	1.83	6.97	1.88
Processing Texts	3.05	1.16	4.06	1.42	6.13	2.38
Reports and Essays	4.03	2.08	5.25	2.05	6.24	2.29
Note Taking	3.75	2.40	4.84	2.17	5.89	2.52
Average of Scales	3.79	1.45	5.12	1.60	6.65	1.54

M=Mean, SD=Standard deviation

Coding scheme for CEFR Categories: 1 (A1); 2 (A2); 3 (A2+); 4 (B1), 5 (B1+); 6 (B2); 7 (B2+); 8 (C1); 9 (C2)

Table 5. ***Descriptive analysis of the raters' assessment of sample students' texts across the PYP levels***

	Elementary n=14		Intermediate n=55		Advanced n=36	
Rating Categories	M	SD	M	SD	M	SD
Range	3.57	1.21	3.90	1.32	5.05	1.28
Coherence	3.50	1.07	3.92	1.35	4.79	1.38
Accuracy	3.47	1.09	3.67	1.26	4.83	1.37
Description	3.55	1.22	3.82	1.28	4.86	1.36
Overall	3.56	1.13	3.87	1.29	4.96	1.28
Average score	3.53	1.14	3.83	1.30	4.88	1.33

M=Mean, SD=Standard deviation

Coding scheme for Manual Grid: 1 (A1); 2 (A2); 3 (A2+); 4 (B1), 5 (B1+); 6 (B2); 7 (B2+); 8 (C1); 9 (C2)

8.4 Appendix 4.

Differences between elementary, intermediate and advanced groups on students' self-assessments

One-way ANOVA was used to identify differences across the PYP levels for the students' assessments. After performing the analysis, Levene's test (Levene 1960) was checked. This test "tests whether the

variance in scores is the same for each of the three groups" (Pallant 2013: 262). Where Levene's test indicated there was no violation of the assumption of homogeneity of variance, ANOVA was used (Table 4); when the assumption of equal variances was violated, the non-parametric analysis of variance (Brown-Forsythe and Welch Tests), as mentioned in Green (2013), were used instead (Table 5).

If the significance (P-value) was <0.05, this indicates a significant difference between the mean scores between the three groups. However, this does not show "which group is different from which other group" (Pallant 2013: 262). For this reason, a post-hoc test, i.e., Tukey's Honestly Significant Difference (HSD) test (Pallant 2013) (if there is no violation to the assumption of homogeneity of variances; Table 6) or Tamhane's T2 (Green 2013) (with heterogeneity of variances; Table 7), were used to check the significance between each pair of the three PYP groups. Post-hoc tests are only utilised if significant differences in means are identified (Pallant 2013: 263).

Table 6 shows the CEFR-based categories for which ANOVA was used.

Table 6. ***One-way analysis of variance of students' self-assessment of CEFR levels across PYP levels***

CEFR Categories	*SS*	*df*	*MS*	*F*	*P-value*	η^2
What Students Can Write						
Between Groups	488.83	2	244.42	52.58	<0.001	0.16
Within Group	2393.82	515	4.65			
Total	2882.65	517				
Reports and Essays						
Between Groups	634.05	2	317.02	60.31	<0.001	0.19
Within Group	2686.11	511	5.26			
Total	3320.16	513				
Note Taking						
Between Groups	279.96	2	139.98	26.89	<0.001	0.095
Within Group	2665.77	512	5.21			
Total	2945.74	514				

SS=Sum of squares, df=degrees of freedom, MS=mean square, F=F ratio, η^2=Effect size M=Mean, SD=Standard deviation, df=degrees of freedom, η^2=Effect size: 0.02=small; 0.13=medium; 0.26=large.

Table 7 shows tests for equality of means for which non-parametric tests were used.

Table 7. ***Robust test of equality of mean of students' self-assessment of their CEFR levels across the three PYP levels***

CEFR Categories	*Statistic*	*df1*	*df2*	*P-value*
Overall Written Production				
Welch	56.05	2	186.89	<0.001
Brown-Forsythe	46.07	2	219.18	<0.001
Overall Written Interaction				
Welch	61.47	2	199.63	<0.001
Brown-Forsythe	69.48	2	338.76	<0.001

CEFR Categories	*Statistic*	*df1*	*df2*	*P-value*
Type of Texts				
Welch	44.49	2	199.82	<0.001
Brown-Forsythe	49.86	2	338.40	<0.001
Vocabulary Range & Control				
Welch	46.06	2	194.25	<0.001
Brown-Forsythe	51.53	2	316.85	<0.001
Grammatical Accuracy				
Welch	13.99	2	188.51	<0.001
Brown-Forsythe	14.90	2	282.66	<0.001
Orthographic Control				
Welch	29.50	2	191.96	<0.001
Brown-Forsythe	25.11	2	242.60	<0.001
Processing Texts				
Welch	52.33	2	205.86	<0.001
Brown-Forsythe	62.06	2	362.55	<0.001

df=degrees of freedom

Table 8 shows the post hoc Tukey honestly significant difference (HSD) test of pairwise differences between groups on student self-assessments.

Table 8. ***Post-hoc Tukey HSD of students' self-assessment of their CEFR levels across the three PYP levels (for items with homogeneity of variances)***

Dependent Variable	(I) PYP levels	(J) PYP levels	Mean difference (I-J)	Std. error	P-value
What Students Can Write	Elementary	Intermediate	-0.48	0.29	0.22
		Advanced	**-2.40***	0.30	<.001
	Intermediate	Advanced	**-1.92***	0.21	<0.001
Reports and Essays	Elementary	Intermediate	-0.36	0.31	0.46
		Advanced	**-2.61***	0.32	<0.001
	Intermediate	Advanced	**-2.25***	0.22	<0.001
Note Taking	Elementary	Intermediate	-0.22	0.30	0.74
		Advanced	**-1.72***	0.32	<0.001
	Intermediate	Advanced	**-1.50***	0.22	<0.001
Conditions and Limitations	Elementary	Intermediate	-0.45	0.44	0.57
		Advanced	**-2.19***	0.46	<0.001
	Intermediate	Advanced	**-1.74***	0.31	<0.001

Table 9 shows the post hoc Tamhane test of pairwise differences between groups on student self-assessments for items with heterogeneity of variances.

Table 9. ***Post hoc Tamhane test (heterogeneity of variances) of students' self-assessment of their CEFR levels across the three PYP levels***

Dependent Variable	(I) PYP levels	(J) PYP levels	Mean difference (I-J)	Std. error	P-value
Overall Written Production	Elementary	Intermediate	-0.67	0.31	0.092
		Advanced	**-2.34***	0.30	<0.001
	Intermediate	Advanced	**-1.67***	0.18	<0.001
Overall Written Interaction	Elementary	Intermediate	-0.30	0.28	0.66
		Advanced	**-2.74***	0.31	<0.001
	Intermediate	Advanced	**-2.44***	0.24	<0.001
Types of Texts the Students can write	Elementary	Intermediate	-0.33	0.28	0.55
		Advanced	**-2.33***	0.31	<0.001
	Intermediate	Advanced	**-2.00***	0.23	<0.001
Vocabulary Range & Control	Elementary	Intermediate	-0.40	0.26	0.34
		Advanced	**-2.30***	0.30	<0.001
	Intermediate	Advanced	**-1.90***	0.22	<0.001
Grammatical Accuracy	Elementary	Intermediate	-0.77	0.35	0.083
		Advanced	**-1.81***	0.38	<0.001
	Intermediate	Advanced	**-1.04***	0.26	<0.001
Orthographic Control	Elementary	Intermediate	-0.36	0.36	0.70
		Advanced	**-1.93***	0.36	<0.001
	Intermediate	Advanced	**-1.57***	0.23	<0.001
Processing Texts	Elementary	Intermediate	**-.58***	0.21	0.020
		Advanced	**-2.31***	0.25	<0.001
	Intermediate	Advanced	**-1.73***	0.20	<0.001

Bold with *=significant results

8.5 Appendix 5

Differences between elementary, intermediate and advanced groups for tutor assessments

One-way ANOVA was used to identify differences across the PYP levels for the tutor assessments. Where there was no violation of the assumption of homogeneity of variance, ANOVA was used (Table 10); when the assumption of equal variances was violated, the non-parametric analysis of variance (Brown-Forsythe and Welch Tests) were used (Table 11). A post-hoc Tukey's HSD (if there is no violation to the assumption of homogeneity of variances; Table 12) or Tamhane's T2 (with heterogeneity of variances; Table 13) were used.

Table 10. ***One-way analysis of variance of tutors' assessment across PYP levels***

CEFR Categories	*SS*	*df*	*MS*	*F*	*P-value*	η^2
Overall written Production						
Between Groups	654.09	2	327.05	84.91	<0.001	0.24
Within Group	2006.79	521	3.85			
Total	2660.88	523				
What Students Can Write						
Between Groups	590.42	2	295.21	80.05	<0.001	0.23
Within Group	1928.81	523	3.69			
Total	2519.22	525				
Reports and Essays						
Between Groups	253.37	2	126.69	27.38	<0.001	0.10
Within Group	2221.15	480	4.63			
Total	2474.52	482				
Note Taking						
Between Groups	250.54	2	125.27	22.78	<0.001	0.08
Within Group	2640.20	480	5.50			
Total	2890.74	482				

SS=Sum of squares, df=degrees of freedom, MS=mean square, F=F ratio, η^2=Effect size: 0.02=small; 0.13=medium; 0.26=large.

Table 11. ***Robust test of equality of mean of tutors' assessment across PYP levels***

CEFR Scales	*Statistic*	*df1*	*df2*	*P-value*
Overall Written Interaction				
Welch	63.84	2	242.69	<0.001
Brown-Forsythe	60.26	2	357.31	<0.001
Type of Texts				
Welch	64.67	2	235.86	<0.001
Brown-Forsythe	59.86	2	317.83	<0.001
Vocabulary Range & Control				
Welch	59.00	2	253.36	<0.001
Brown-Forsythe	60.82	2	426.13	<0.001
Grammatical Accuracy				
Welch	40.63	2	233.06	<0.001
Brown-Forsythe	44.31	2	366.37	<0.001
Orthographic Control				
Welch	77.17	2	159.67	<0.001
Brown-Forsythe	63.16	2	166.48	<0.001
Processing Texts				
Welch	94.79	2	194.58	<0.001
Brown-Forsythe	116.16	2	357.74	<0.001

df=degrees of freedom

Table 12. ***Tukey HSD of tutors' assessment across the three PYP levels***

Dependent Variable	(I) PYP levels	(J) PYP levels	Mean difference (I-J)	Std. error	P-value
Overall Written Production	Elementary	Intermediate	**-1.61***	0.24	<0.001
		Advanced	**-3.18***	0.25	<0.001
	Intermediate	Advanced	**-1.57***	0.19	<0.001
Types of Texts Students can write	Elementary	Intermediate	**-1.35***	0.26	<0.001
		Advanced	**-2.82***	0.27	<0.001
	Intermediate	Advanced	**-1.48***	0.20	<0.001
Reports and Essays	Elementary	Intermediate	**-1.22***	0.31	<0.001
		Advanced	**-2.21***	0.31	<0.001
	Intermediate	Advanced	**-0.99***	0.21	<0.001
Note Taking	Elementary	Intermediate	**-1.10***	0.33	<0.001
		Advanced	**-2.15***	0.34	<0.001
	Intermediate	Advanced	**-1.05***	0.23	<0.001
Average of all scales	Elementary	Intermediate	**-1.33***	0.19	<0.001
		Advanced	**-2.86***	0.20	<0.001
	Intermediate	Advanced	**-1.53***	0.15	<0.001

Table 13. ***Post hoc Tamhane of tutors' assessment across the three PYP levels***

Dependent Variable	(I) PYP levels	(J) PYP levels	Mean difference (I-J)	Std. error	P-value	95% Confidence interval	
						Lower bound	Upper bound
Overall Written Interaction	Elementary	Intermediate	**-1.51***	0.25	<0.001	-2.12	-.89
		Advanced	**-2.76***	0.25	<0.001	-3.37	-2.15
	Intermediate	Advanced	**-1.26***	0.19	<0.001	-1.71	-.80
What students Can Write	Elementary	Intermediate	**-1.46***	0.23	<0.001	-2.01	-.91
		Advanced	**-2.99***	0.25	<0.001	-3.58	-2.40
	Intermediate	Advanced	**-1.54***	0.19	<0.001	-1.99	-1.08
Vocabulary Range and Control	Elementary	Intermediate	**-1.16***	0.20	<0.001	-1.65	-.67
		Advanced	**-2.52***	0.23	<0.001	-3.06	-1.95
	Intermediate	Advanced	**-1.35***	0.20	<0.001	-1.82	-.88
Grammatical Accuracy	Elementary	Intermediate	**-1.11***	0.23	<0.001	-1.66	-.56
		Advanced	**-2.29***	0.26	<0.001	-2.91	-1.67
	Intermediate	Advanced	**-1.18***	0.20	<0.001	-1.65	-.71
Orthographic Control	Elementary	Intermediate	**-0.67**	0.34	0.147	-1.50	.16
		Advanced	**-2.75***	0.35	<0.001	-3.59	-1.91
	Intermediate	Advanced	**-2.08***	0.18	<0.001	-2.51	-1.65
Processing Texts	Elementary	Intermediate	**-1.51***	0.25	<0.001	-1.43	-.59
		Advanced	**-2.76***	0.25	<0.001	-3.63	-2.54
	Intermediate	Advanced	**-1.26***	0.19	<0.001	-2.54	-1.60

Bold with *=significant results

8.6 Appendix 6

Differences between elementary, intermediate and advanced groups on rater assessments

Table 14 shows the ANOVA for differences across the PYP levels for the rater assessments and Table 15 shows the post-hoc Tukey's HSD.

Table 14. ***One way ANOVA of raters' assessment across PYP levels***

		SS	*df*	*MS*	*F*	*P-value*	η^2
	Between Groups	37.28	2	18.64			
Range	Within Groups	66.32	99	0.67	27.823	*p<0.001*	0.36
	Total	103.59	101				
	Between Groups	24.64	2	12.32			
Coherence	Within Groups	65.04	99	0.66	18.76	*p<0.001*	0.28
	Total	89.7	101				
	Between Groups	35.33	2	17.66			
Accuracy	Within Groups	60.32	99	0.61	28.99	*p<0.001*	0.37
	Total	95.65	101				
	Between Groups	29.93	2	14.97			
Description	Within Groups	61.04	99	0.62	24.28	*p<0.001*	0.33
	Total	90.97	101				
	Between Groups	33.23	2	16.61			
Overall	Within Groups	64.11	99	0.65	25.66	*p<0.001*	0.34
	Total	97.34	101				

SS=Sum of squares, df=degrees of freedom, MS=mean square, F=F ratio, η^2=Effect size: 0.02=small; 0.13=medium; 0.26=large.

Table 15. ***Post hoc Tukey analysis of range, coherence, accuracy, description, and overall grouped by PYP levels***

Dependent Variable	(I) PYP Levels	(J) PYP Levels	Mean Difference (I-J)	Std. Error	P-value	95% Confidence Interval	
						Lower Bound	Upper Bound
Range	Elementary	Intermediate	-0.33	0.25	0.40	-0.93	0.27
		Advanced	-1.53	0.27	**<0.001***	-2.15	-0.89
	Intermediate	Advanced	-1.19	0.18	**<0.001***	-1.61	-0.77
Coherence	Elementary	Intermediate	-0.42	0.25	0.227	-1.01	0.18
		Advanced	-1.33	0.26	**<0.001***	-1.97	-0.70
	Intermediate	Advanced	-.92	0.18	**<0.001***	-1.33	-0.50
Accuracy	Elementary	Intermediate	-0.20	0.24	0.674	-0.78	0.37
		Advanced	-1.40	0.25	**<0.001***	-2.00	-0.79
	Intermediate	Advanced	-1.19	0.17	**<0.001***	-1.60	-0.79
Description	Elementary	Intermediate	-0.27	0.24	.0507	-0.85	0.31
		Advanced	-1.34	0.26	**<0.001***	-1.95	-0.74
	Intermediate	Advanced	-1.07	0.17	**<0.001***	-1.48	-0.67
Overall	Elementary	Intermediate	-0.31	0.25	0.423	-0.90	0.28
		Advanced	-1.44	0.26	**<0.001***	-2.06	-0.81
	Intermediate	Advanced	-1.12	0.17	**<0.001***	-1.54	-0.71

Bold with *=significant results

8.7 Appendix 7

RQ2: Student versus teachers paired t-test and correlation

Table 16 shows the paired t-test between students and teachers for each scale, separated by PYP level.

Table 16. ***Paired differences between self-and tutors' assessment in each PYP level***

CEFR Scales	PYP students		PYP tutors					Cohen's
	M	SD	M	SD	t	df	*P*	d
Elementary (n=72)								
Overall Written Production	5.62	2.33	4.41	1.92	3.72	70	**<0.001**	0.44
Overall Written Interaction	3.96	2.10	4.11	2.12	-0.56	70	0.576	-0.07
Type of Texts	3.94	2.06	4.46	2.38	-1.54	70	0.128	-0.18
What Can They Write	4.40	2.26	3.47	1.85	3.07	71	**0.003**	0.36
Vocabulary Range & Control	3.59	2.00	3.86	1.66	-0.91	70	0.367	-0.11
Grammatical Accuracy	4.34	2.70	3.85	1.95	1.41	70	0.164	0.17
Orthographic Control	4.78	2.92	4.24	2.59	1.04	50	0.304	0.15
Processing Texts	3.80	1.61	3.00	1.22	2.84	50	**0.006**	0.40
Reports and Essays	4.00	2.62	4.10	2.05	-0.22	49	0.826	-0.03
Note Taking	5.04	2.69	3.80	2.46	2.70	50	**0.009**	0.38
Average Scales	4.49	1.59	3.97	1.65	2.24	71	**0.028**	0.26
Intermediate (n=232)								
Overall Written Production	6.26	2.17	5.97	2.08	1.52	226	0.129	0.10
Overall Written Interaction	4.23	2.33	5.60	2.17	-6.61	226	**<0.001**	-0.44
Type of Texts	4.28	2.26	5.79	2.25	-7.77	228	**<0.001**	-0.51
What Can They Write	4.78	2.25	4.94	1.86	-0.85	228	0.394	-0.06
Vocabulary Range & Control	3.87	1.95	4.94	1.86	-6.64	230	**<0.001**	-0.44
Grammatical Accuracy	5.05	2.37	4.95	1.75	0.57	230	0.570	0.04
Orthographic Control	5.47	2.70	4.87	1.86	2.88	217	**0.004**	0.19
Processing Texts	4.36	1.70	4.01	1.41	2.48	217	2.014	0.17
Reports and Essays	4.55	2.31	5.20	2.10	-3.09	210	**0.002**	-0.21
Note Taking	5.43	2.18	4.81	2.20	3.00	211	**0.003**	0.21
Average Scales	4.89	1.51	5.18	1.67	-2.15	230	**0.032**	-0.14
Advanced (n=170)								
Overall Written Production	7.96	1.65	7.62	1.82	1.87	168	0.064	0.14
Overall Written Interaction	6.74	2.56	6.90	1.83	-0.66	168	0.510	-0.05
Type of Texts	6.35	2.47	7.24	1.81	-4.01	169	**<0.001**	-0.31
What Can They Write	6.86	1.95	6.56	2.05	1.45	169	0.150	0.11
Vocabulary Range & Control	5.86	2.40	6.31	2.19	-1.75	168	0.082	-0.13
Grammatical Accuracy	6.14	2.89	6.19	2.19	-0.19	168	0.847	-0.01
Orthographic Control	6.99	2.16	7.05	1.78	-0.30	168	0.762	-0.02
Processing Texts	6.12	2.22	6.24	2.34	-0.51	168	0.613	-0.04
Reports and Essays	6.78	2.07	6.26	2.26	2.18	169	**0.030**	0.17
Note Taking	6.90	2.19	5.96	2.43	3.84	168	**<0.001**	0.30

M= Mean, SD=Standard deviation

Coding scheme for CERF Scales: 1 (A1); 2 (A2); 3 (A2+); 4 (B1), 5 (B1+); 6 (B2); 7 (B2+); 8 (C1); 9 (C2)

Cohen's d_z calculated as Mean misalignment/SD of misalignment. Cohen's d calculated as 2 x t/sqrt, 0.2=small effect; 0.5=medium; 0.8=large

Bold = significant result

Table 17 shows the correlation between students and teachers' scores, the weighted kappa (measure of agreement) and the percentages of scores with exact agreement (identical level assigned), or agreements within one or two levels.

Table 17. ***Correlation and agreement between ratings of self- and tutors' assessment***

CEFR Scales	Correlation (r) (n=517)	Weighted Kappa (n=517)	% exact agreement	% within one adjacent CEFR level	% within two adjacent CEFR levels
Overall Written Production	0.29 P<0.001	0.27	31.5	38.9	65.5
Overall Written Interaction	0.22 P<0.001	0.22	23.3	33.2	62.7
Types of Texts the Students can write	0.29 P<0.001	0.25	23.6	31.5	60.4
What Students can write	0.28 P<0.001	0.28	25.7	31.6	67.9
Vocabulary Range and Control	0.25 P<0.001	0.25	21.7	35.2	61.6
Grammatical Accuracy	0.23 P<0.001	0.19	15.9	40.8	61.8
Orthographic Control	0.26 P<0.001	0.26	21.5	31.3	68.0
Processing Texts	0.30 P<0.001	0.32	29.9	48.4	73.7
Reports and Essays	0.23 P<0.001	0.15	20.2	45.9	65.0
Note Taking	0.18 P<0.001	0.15	22.7	39.4	59.5

8.8 Appendix 8.

RQ2. Comparisons of students, teachers and raters' assessments

Table 18 and 19 show the Tukey's post hoc tests, firstly (Table 16) with data for all students across the PYP levels and secondly (Table 17) separated by PYP level.

Table 18. ***Tukeys post hoc analysis for scores grouped as to the type of raters***

(I) Type	(J) Type					
	Students' self-assessment		Teachers' assessment		Raters' assessment	
	Mean Difference (I-J)	p-value	Mean Difference (I-J)	p-value	Mean Difference (I-J)	p-value
Students' self-assessment			-0.031	0.99	1.28*	<0.001
Teachers' assessment	0.031	0.99			1.31*	<0.001
Raters' assessment	-1.28*	<.001	-1.31*	<0.001		

* The mean difference is significant at the 0.05 level.

Table 19. ***Post hoc Tukey analysis of PYP level grouped by assessor and level***

PYP Levels	(I) Type	(J) Type	Mean Difference (I-J)	p-value
Elementary	Self	Tutors	-0.03	0.998
		Raters	1.44*	0.004
	Tutors	Raters	1.47*	0.004
Intermediate	Self	Tutors	-0.22	0.67
		Raters	0.92*	0.001
	Tutors	Raters	1.14*	<0.001
Advanced	Self	Tutors	0.26	0.67
		Raters	1.77*	<0.001
	Tutors	Raters	1.52*	<0.001

* The mean difference is significant at the 0.05 level.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Application of the CEFR to an Arabic Corpus: A Case Study

Aziza Zaher, Durham University

https://doi.org/10.37546/JALTSIG.CEFR7-5

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment *(CEFR) was developed by the Council of Europe and first published in 2001. It has since evolved significantly and new volumes have been published; most recently, the CEFR Companion Volume (CEFR/CV) in 2020. The CEFR aims to provide the basis for L2 learning, teaching, and assessment of European languages. However, it has been widely used around the world in non-European contexts.*

This article presents a case study of the application of the CEFR to an Arabic corpus comprising 214 texts produced by first year students at Zayed University in the UAE, which is part of a bilingual corpus in Arabic and English. This article focuses on the application of the CEFR to the Arabic texts which posed specific challenges, including Arabic diglossia whereby there are two distinct varieties of the language used for writing and speaking. Furthermore, the complexities of Arabic grammar include that it has formal features which only appear in writing. There is also some overlap between Arabic and other languages, particularly English, as many English expressions are used in everyday life in Arab societies. These factors, among others, lead to unique issues to consider when applying the CEFR to a written Arabic corpus. However, due to the generic nature of the CEFR descriptors, they have been applied successfully to the assessment of the Arabic written corpus, which provides the basis for further applications of the CEFR to other competencies in Arabic and to other non-European languages. This article describes the process of rating the corpus, outlines the practical implications of the application of the CEFR to an Arabic written corpus and presents an overview of student performance mapped across the six CEFR levels.

Keywords: CEFR, Arabic, written Corpus, Assessment, Non-European languages, Diglossia

1 Introduction to the CEFR

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) was published in 2001 as the culmination of a lengthy process that aimed to support communicative language learning and teaching across Europe. The CEFR has various political, socio-cultural, and educational aims, and was envisaged as a tool to help language planners, educators, and learners in course design, assessment, and certification across Europe and beyond: "It aims to facilitate transparency and coherence between the curriculum, teaching and assessment within an institution and transparency and coherence between institutions, educational sectors, regions and countries" (Council of Europe 2020: 27). The CEFR was perceived as a flexible document that can be used by practitioners in different ways. Moreover, it has been evolving, with the *CEFR Companion Volume with New Descriptors* published in 2018, and *the CEFR Companion Volume* published in 2020, and myriad other relevant resources which are available on the CEFR website, and beyond. The Companion Volume updates the original framework by adding descriptors for online interaction, collaborative learning, and mediating text, as well as descriptors for plurilingualism and pluriculturalism, and a chapter on sign language scales and descriptors (Council of Europe 2020).

The CEFR emphasizes fundamental concepts, such as the role of the learner as a social agent, and the co-construction of meaning in interaction, as well as the notions of mediation, and plurilingual/pluricultural competences. It provides a comprehensive descriptive scheme for language proficiency across Common Reference Levels from A1 to C2, and it is based on 'can-do' statements that provide a clear yet nuanced instrument for the assessment of progress and proficiency. The CEFR views language as "a vehicle for opportunity and success in the social, educational and professional domain" (Council of Europe 2020: 27), and its 'action-oriented' model guided by the 'can-do' statements focuses on real-life tasks and the learner's proficiency rather than their deficiency.

The main purpose of the CEFR is to improve the quality and effectiveness of language learning and teaching. It has been argued that the CEFR project has never been about assessment or harmonisation, but rather about learning and teaching (North et al. 2022: 27); however, the CEFR has key applications in both assessment and accreditation. The CEFR aims to promote co-operation between educational institutions in different countries, provide a basis for the mutual recognition of language qualifications; and assist learners, teachers, and course designers among others to co-ordinate their efforts. This is achieved via common reference levels and illustrative descriptors which provide a metalanguage for language professionals to facilitate communication, networking, mobility, and recognition of qualifications (Council of Europe 2001). The CEFR has been a flexible tool used for many purposes and in various contexts, both European and non-European. This has been its purpose from the beginning, as the CEFR does not set out to tell practitioners what to do, or how to do it, as it raises questions, rather than answering them. Moreover, the CEFR is innovative in its approach; North emphasizes that "the main purpose of the CEFR project is to stimulate innovation in language education through the concepts of the user/learner acting as a social agent, (co)constructing meaning and knowledge, while drawing on their full plurilingual repertoire to do so" (2022: 1). However, the CEFR has faced many criticisms which extend from its theoretical basis to its interpretations and applications (Alderson 2007; Deysgers 2019; Hulstijn 2007). Its scales have been criticized for being underspecified (Neff-van Aertselaer 2013) and impressionistic in their wording (Alderson 2007), and some of the scales read as outdated or Eurocentric which may limit their applicability to the global community of language learners (Foley 2019). Nevertheless, there have been attempts to apply the CEFR outside Europe in relation to teaching English as a foreign language, e.g. in China, Japan, Turkey, among others (Hazar 2021; Lu 2017; Negishi 2012; O'Dwyer 2017). For example, countries in the ASEAN region adapted the CEFR for the teaching and assessment of English as a foreign language in their contexts, which resulted in different version of the framework, e.g. CEFR-J for Japan, CEFR-M for Malaysia, CEFR-V for Vietnam, and the CCFR or the Common Chinese Framework of Reference for Languages. Each of these versions reflects the local context, needs of learners, and the educational systems of the country in which it was developed (Foley 2019). Additionally, there have also been limited attempts to adapt and apply the CEFR to non-European languages. This paper describes how it has been used in the assessment of an Arabic written corpus. The following sections introduce the Arabic language and the relationship between the CEFR and Arabic.

2 The Arabic Language

Arabic is the official language of 22 countries and the native language of over 400 million speakers in North Africa and Western Asia. It belongs to a group of languages known as the Semitic languages (Versteegh 2001), which in turn belong to a broader group of languages, termed Afro-Asiatic (Ryding 2005). This distinguishes it from many European languages which belong to a family of languages known as the Indo-European.

"The linguistic situation in the Arab world is strongly characterised by diglossia" (Ryding 1991: 212). The term *diglossia* was first used by Marçais (1930), but it received a lot of attention with Ferguson's seminal paper in 1959 in which he describes the situation in which two dialects or varieties exist, one which may be a vernacular or spoken dialect alongside a standard written or formal variety (Ferguson 1959; Horn 2015; Kaye 2001). Kaye (2001) argues that colloquial Arabic is grammatically and lexically less complex

than Modern Standard Arabic (MSA) and that there is a continuum between MSA and colloquial forms as well as between one colloquial dialect and another to the point that some uneducated people may find MSA unintelligible, and a speaker of a certain dialect may find another one unintelligible.

Badawi (1973) argues that MSA and colloquial dialects are independent varieties of the same language, each with its own lexicon and grammar, and they differ in the context of use. However, there is no clear-cut division between standard Arabic and colloquial dialects. Instead, there is a great deal of overlap and there are various geographic and socio-cultural variations or levels that exist within the language; sometimes they are quite distinct and at other times they are very subtle and hard to notice.

Arabic native speakers learn their local spoken dialect as their mother tongue, and then they learn MSA at school. It is then that the child becomes diglossic. So, the experience of learning MSA is like the experience of learning a second language (L2), especially given that MSA is nobody's mother tongue (Maamouri 1998). It should be noted, however, that MSA has higher prestige than spoken dialects even though they are used in different contexts and although they have distinct lexical and grammatical inventories, since MSA must be learned and is associated with having received an education. It would be inappropriate to use MSA in everyday life, but it is equally inappropriate to use spoken lexis in an academic essay. The two varieties co-exist in literary output that involves dialogue in spoken Arabic and narration in MSA.

Although Arabic speakers learn MSA at school, they are exposed to it much more than they are to a second language since mainstream media have strong elements of MSA; for example, in news, documentaries, dubbed TV shows, and so on. Moreover, a considerable proportion of the cartoons that children watch is also dubbed in MSA. Consequently, Arabic speakers are exposed to MSA at a young age, but they do not effectively produce it until they go to school, except in very limited contexts such as performing religious duties. Therefore, the experience of Arabic native speakers with MSA is somewhat similar to their experience with a second language. Moreover, there are substantial similarities between the native dialects and MSA, which makes the learning of MSA easier than learning a foreign language. This complex relationship between spoken and written Arabic is one of the major challenges faced in the application of the CEFR to Arabic.

3 Arabic and the CEFR

It has been noted that there is an increasing familiarity with the CEFR terminology and scales outside the EU, including the Arab world, yet there is no systematic effort to apply the CEFR to Arabic, however, there are some sporadic attempts. In 2021, an official Arabic version of the *CEFR Companion Volume* has been published, which should have an impact on Arabic language teaching, assessment, and research. However, it has also been noted that there is no coherent agenda for the application of the CEFR or a similar framework for Arabic teaching (Soliman 2018: 122). Soliman discusses the difficulties faced in the design of detailed CEFR level descriptors for Arabic in the light of the vast differences between Arabic and European languages, e.g. Arabic diglossia, the reality of language learning and use, and the linguistic complexity of Arabic. Therefore, the application of the CEFR to Arabic has been attempted on an individual or very small-scale basis, and mostly in an unsystematic way.

The CEFR has been mainly applied to assessment in Arabic, and there are many Arabic language tests which claim to be aligned with the CEFR. Soliman (2018: 213) lists some of these tests, e.g. the American Council on the Teaching of Foreign Languages (ACTFL), the Al-Arabiyya Test developed by Eckehard Schulz, the Standardised Arabic Test developed by the Saudi Electronic University, the ILA certificate in Arabic, and the TELC Arabic language test, among others.

Moreover, many Arabic courses claim to be aligned with the CEFR. This is common in courses taught by language centres in many UK universities. This could be influenced by the way European languages are described in these institutions, so that the same terminology is used with Arabic courses. Moreover, some Arabic qualifications and resources also claim to be mapped against the CEFR.

The CEFR has also impacted research on Arabic pedagogy. There are academic papers and conference presentations that deal with the application of the CEFR to Arabic in different contexts (e.g., Al-Jarf & Mingazova 2020; Mohamed 2021, Mohamed 2023; Soliman 2018). This paper aims to contribute to this body of research by discussing the application of the CEFR to a corpus of Arabic texts that were produced as part of a bilingual learners' corpus.

4 Introduction to ZAEBUC

The Zayed Arabic-English Bilingual Undergraduate Corpus (ZAEBUC) is an annotated Arabic-English bilingual writer corpus comprising short essays by first-year university students at Zayed University in the United Arab Emirates. "The corpus comprises short essays written by 397 first-year university students totalling 388 English essays (87.6K words) and 214 Arabic essays (33.3K words)" (Habash & Palfreyman 2022: 79). It is available in both raw and corrected versions and is an open resource available for researchers. Moreover, it has been rated using the CEFR. Although the corpus is bilingual, this chapter focuses on the assessment of the Arabic texts using the CEFR. The assessment process is described as well as the challenges faced in the application of the CEFR to Arabic, and then a commentary is provided on the outcome of the assessment of the Arabic texts using the CEFR, and the potential of furthering the application of the CEFR or a similar framework to Arabic.

5 Application of the CEFR to ZAEBUC

5.1 The Rating Process

The assessment of the ZAEBUC using the CEFR involved several steps, as the raters were applying the CEFR to a corpus of written Arabic texts for the first time. Initially, two raters discussed the potential and limitations of applying the CEFR to both the Arabic and English samples in the corpus. They worked on 10 Arabic and English texts written by the same student. A subsequent meeting discussed the outcomes of the assessments. It became clear that the assessments may have been influenced by the fact that the raters had access to English and Arabic samples by the same writer. Moreover, the raters focused on different aspects of writers' performance. For example, for one rater accuracy seemed crucial, while for the other the range of lexis and the cohesion of texts were deemed more important than accuracy, and the rater was more tolerant towards accepting grammar errors if the range of lexis was wider.

Then,10 randomized samples, where the raters did not get the English and Arabic texts by the same student, were assessed, followed by another meeting in which it was decided to randomize the samples before assessing the corpus. These discussions helped the raters to consider which criteria were important to each of them and to agree on common grounds. Based on the CEFR, criteria selected at this stage involved writer's ability to address the topic in a clear, organized way, the range of lexis and structures used, the use of cohesive devices, the accuracy of grammatical structures and the appropriateness of lexical choices.

Since both raters were native speakers of Arabic with excellent knowledge of English and considerable teaching experience, the decision was made to involve a third rater who was a native speaker of English with teaching experience and excellent knowledge of Arabic. There was another round of assessment of samples in both languages by the three raters, followed by a meeting with an expert on the CEFR who discussed divergences in the assessments and assisted in normalizing a sample of English texts. After that meeting, the three raters completed their independent assessments of the corpus and entered their ratings on Google forms.

Most of the initial discussions focused on the assessment of Arabic samples, as the CEFR was designed for European languages and the raters wanted to ascertain its applicability to Arabic; especially that Arabic was L1 of the students and the CEFR was designed for L2 contexts. The CEFR proved to be applicable to the Arabic samples. Due to the generic nature of the CEFR descriptors, it was possible to apply them

to Arabic L1 samples. However, as pointed out by Neff-van Aertselaer (2013: 200), the "reference-level descriptors for each of the 6 broad competence bands are under-specified". This under-specification led to some divergence in interpretations of the criteria, and different raters relied on their backgrounds in interpreting the criteria and applying them to the samples. As a result, there were differences between the ratings and the average of the three assessments was used as the final assessment for each text.

5.2 Agreement between Raters

There has been an acceptable level of agreement between the raters. In 28.30% of the samples, all three raters gave the same assessment. In 90.57% of the samples, at least two raters gave the same assessment. Where raters differed, there was one band difference between the raters in 54.72% of the cases, and in 20.28% of cases the difference was two bands.

Based on their experience with the ZAEBUC, the raters found the CEFR to be an effective tool for the assessment of the writing competence in Arabic, and the participants' performance could be mapped across the scales of the CEFR. The scales were deemed very appropriate in assessing the samples, as they provided the raters with a flexible, consistent, and reliable tool for the assessment of competence. The CEFR descriptors could be applied consistently across the corpus. The standardization meetings showed that the assessors might have placed different weights on certain aspects of the participants' performance, but generally there was agreement as to what constituted A-, B- or C-level performance in a piece of writing. However, within the same scale, the same text was sometimes assessed as level 1 or 2 depending on the experience and focus of different raters, but it was not common for raters to assess the same text for different scales. Comprehensibility, range of lexis and structures, coherence and cohesion, thematic development, and accuracy were considered crucial criteria for all three assessors. However, the type of errors and their significance were sometimes debated among raters, such as the importance of certain formal grammatical features which did not significantly affect the meaning even if they were not used accurately.

5.3 Students' Scores

The scores of assessing the Arabic samples ranged from A2 to C1. 3% of the samples achieved A2. This means that the students who were assessed as A2 could "produce simple texts on familiar subjects of interest, linking sentences with connectors like 'and', 'because' or 'then.'" Most of the samples were in the B scale. 52% of the samples achieved B1, which means that these students could "produce a text on a topical subject of personal interest, using simple language to list advantages and disadvantages, and give and justify their opinion." Of the samples, 38% of them achieved B2, which means that those students could "produce an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail." Additionally, 5% of the samples achieved C1, which means that the students could "produce clear, well-structured expositions of complex subjects, underlining the relevant salient issues". None of the samples was assessed as C2, and this could be attributed to the nature of the task, since students were not required to "set out multiple perspectives on complex academic or professional topics, clearly distinguishing their own ideas and opinions from those in the sources" (Council of Europe 2020: 68).

5.4 Examples from the Arabic Corpus

The broad and generic nature of the descriptors allowed them to be applied to the texts readily. However, there were certain issues pertaining to the Arabic texts that merited extensive discussion in the assessments; for example, the diglossic nature of Arabic. Academic writing was considered a formal activity, and therefore Modern Standard Arabic (MSA) was expected to be used in the Arabic samples;

therefore, deviations from it would be seen as problematic in terms of using the correct register and style. Nevertheless, there were instances of students using colloquial Arabic in their essays.

The raters discussed whether this could be regarded as evidence of plurilingualism, because students might exploit their plurilingual repertoires by using features from their colloquial dialects. It was believed that students used colloquial Arabic because they lacked mastery in the appropriate variety and did not have the competence required to complete the communicative task in MSA as would be expected. As a result, they resorted to compensating, which is "a strategy for maintaining communication when one cannot think of the appropriate expression" (Council of Europe 2020: 69).

The participants' writing samples showed influences from colloquial dialects at the phonological, lexical, syntactic, and stylistic levels. Examples of colloquial features that students used in their writing are presented below. It should be noted that these examples are not based on the entire corpus, but on a randomly selected sample of 10 texts.

At the phonological level, it was possible to notice influences from the phonology colloquial Arabic in students' writing, for example, replacing the sounds /ḍ/ with /ẓ/, replacing the final (*tā' marbūṭah*) with (*tā' maftūḥah*) in certain structures, replacing short vowels with long vowels, and reducing the glottal stop to a short vowel. Table 1 show examples of influences from colloquial Arabic at the phonological level.

Table 1. ***Examples colloquial influences at the phonological and orthographic levels***

(a) replacing the sound *ḍ* with *ẓ*

Error	Correct form
محاظرات *muḥā/ẓa/rāt* 'lectures'	محاضرات *muḥā/ḍa/rāt* 'lectures'
حظاري *ḥa/ẓā/rī* 'civilised'	حضاري *ḥa/ḍā/rī* 'civilised'

(b) replacing the final (*tā' marbūṭah*) with (*tā' maftūḥah*). This change can only be noticed in the orthography of Arabic words on the final syllabus, as the English transcription shows the same sounds.

Error	Correct form
خاصتاً *khāṣtan* especially	خاصةً *khāṣtan* especially
شهرتاً *Shuhratan* fame	شهرةً *Shuhratan* fame
معرفت *ma'rifat* knowledge	معرفة *ma'rifat* knowledge

(c) replacing short vowels with long vowels. This change can only be noticed in the orthography of Arabic words.

Error	Correct form
لهاذا *lihādhā* thus	لهذا *lihādhā* thus
هاذي *Hādhi* this (f.)	هذه *hādhihi* this (f.)

(d) reducing the glottal stop to a short vowel.

Error	Correct form
شي */shay/* thing	شيء */shay'/* thing
نبدي *Neb/dī/* *we start*	نبدأ *Nab/da'/* *we start*

At the lexical level, some students replaced certain lexical items from MAS with their colloquial counterparts. As noted earlier, it would be inappropriate to use colloquial words in an academic essay as required in the task. Thus, students compensated for not knowing the formal words appropriate for the context of academic writing by using the colloquial words they were familiar with, as shown in Table 2.

Table 2. ***Examples of colloquial influences at the lexical level***

(e) replacing certain lexical items from MAS with their colloquial counterpart.

Error	Correct form
زعل *za'al* sorrow	حزن *ḥuzn* sorrow
الأشاعات *al'shā'āt* rumours	الشائعات *ash-shā'i'āt* rumours
اجاوب 'ujāwib I reply	أجيب 'ujīb I reply

At the syntactic level, it has been noted that, very often, the syntactic complexity of standard Arabic is not upheld. There are highly formal features of Arabic grammar that only appear in writing and that are often found difficult to apply by most Arabic speakers as they are not used in spoken dialects, for example the case marking system. This system involves selecting certain endings for words to mark their case, i.e., their function or position in the sentence. Errors in cases do not usually affect the comprehensibility of the text, as it could still be understood correctly despite being grammatically incorrect. This is a typical example of the influence of colloquial Arabic on the writing of students or of confusion about

the appropriate syntax of MSA. In both examples below, case marking rules were either confused when using the nominative case instead of the genitive in the first example or ignored such as in missing the accusative ending in the second example. Case marking is a very formal characteristic of Arabic which is hardly reflected in spoken dialects, and therefore, confusion here could be a feature of the influence of colloquial dialects. Examples of colloquial influences at the syntactic level are shown in Table 3.

Table 3. ***Examples of colloquial influences at the syntactic level***

(f) confusing cases; using nominative instead of accusative and genitive.

Error	Correct form
فأصبح سهل للمجرمون *fa'ṣbaḥa sahl [NOM.SG] lilmujrimūn [NOM.PL]* it became easy for criminals	فأصبح سهلاً للمجرمين *fa'ṣbaḥa sahlan [ACC.SG] lilmujrimīn [GEN.PL]* it became easy for criminals

(g) ignoring cases; not adding the accusative case ending.

سوف يجعله شخص كسول *saūfa yaj'aluhu shakhṣ [NOM.SG] kasūl [NOM.SG]* it will make him a lazy person	سوف يجعله شخصًا كسولاً *sawfa yaj'aluhu shakhṣan [ACC.SG] kasūlan [ACC.SG]* it will make him a lazy person

At the stylistic level, some colloquial expressions were used in the students' writing. These expressions are characteristic of spoken Arabic and would be inappropriate to use in writing, as shown in Table 4.

Table 4. ***Examples of colloquial influences at the stylistic level***

(h) using colloquial expressions instead of standard ones.

بالاخير *bil'ālākhīr* eventually	في نهاية المطاف *fī nihāyat al-maṭāf* eventually
الكلام الفاضي *al-kalām al-fāḍī* trivialities	التفاهات *at-tafāhāt* trivialities

It is interesting to note these instances of colloquial influences in students' writings since their main training in Arabic writing will have been in MSA. But the participants are young people who often communicate with each other through digital and social media. The writing codes typically used in these media are often informal and inconsistent, with a great deal of codeswitching. This inevitably affects the quality of the writing, and the registers students use, especially because a great deal of the language they use is influenced by the features of language used on social media. Bies et al. (2014: 93) noted that "the language used in social media expresses many differences from other written genres: its vocabulary is informal with intentional deviations from standard orthography such as repeated letters for emphasis; typos and non-standard abbreviations are common; and non-linguistic content is written out, such as laughter, sound representations, and emoticons."

Another characteristic of the Arabic texts is their short length, whereas it was noted that the English texts in the corpus were longer. This could partly be due to the nature of Arabic as a "morphologically rich and complex language. Arabic words are agglutinated words, composed by an inflected word form (base) and attachable clitics" (Mallek et al. 2017: 299). However, this only explains some of the differences

found. The average length of the English texts of the corpus was 226 words, in comparison to 155 words for those in Arabic. Some of the Arabic texts scored A2 due to being too short for the topic to be developed properly. Moreover, some Arabic texts were too short to assess at all and were, thus, marked as 'unassessable.' It is interesting to note that some of the texts were so short that they sounded more like tweets.

The Arabic texts also included examples of the use of non-Arabic words, especially for names of places and brands such as Dubai Mall, Expo 2020 and Ferrari. It is worth noting that UAE is a largely bilingual setting, e.g. English and Arabic are used together on all road signs, shop banners and other places, which makes it a norm to blend English terms into Arabic vocabulary. Moreover, many English words are used in everyday life in Arab societies, e.g. in greetings and informal settings. In academic institutions, it is common to find students switching between both Arabic and English while speaking, as in most of the universities, the medium of instruction is English.

6 Discussion

The application of the CEFR to the Arabic learners' corpus has proven to be possible, and the CEFR descriptors lent themselves quite well to the corpus although it contained L1 texts in a non-European language. This is due to the generic nature of the descriptors and the nature of the learning of Arabic which is similar to L2 learning as a result of Arabic diglossia. Although the present study is a small-scale project, it demonstrates the advantage of a framework like the CEFR for the learning, teaching, and assessment of Arabic.

The Arab world is vast, and it includes 22 countries with different regional dialects, educational systems, as well as varying economic and socio-cultural contexts, and thus, such a framework would be extremely useful for the purposes of cooperation and mutual accreditation. There are many challenges that would be faced in the establishment of such a framework, but it is believed to be a very worthwhile endeavour with benefits that can extend beyond Arab nations.

Despite criticisms of the CEFR, it has proven to be effective when adapted to different contexts, such as in the ASEAN region (Foley 2019). These adaptations made it possible to apply the CEFR flexibly in the specific contexts of different countries to achieve a range of purposes. However, it has been noted that many language professionals outside the EU are not familiar with the underlying concepts of CEFR, which may lead many teachers to associate it with testing only, which is a limited view of what the CEFR is about. The application of the CEFR in the ASEAN region has led to the identification of many issues with the educational systems of these countries, e.g. teachers' proficiency in English and understanding of the CEFR, lack of local experts on the CEFR, lack of training on the CEFR, and the limited view of the CEFR as a testing tool, among others (Foley 2019).

Although the CEFR has been used mainly for European languages, even in non-European contexts, there is no reason why it should not be adapted to the context of Arabic. In order to achieve this objective, it will be necessary for language professionals from across the Arab nations to collaborate in a concerted manner. Then, it may be possible, given the existence of the Arab League and its Arab Organisation for Education, Culture and Science, as well as organisations that support the learning and teaching of Arabic such as the Qatar Foundation and many universities and research institutions in the Arab world.

The application of the CEFR for the assessment of Arabic written texts in the current study has shown important findings: the average proficiency of first year university students in Arabic, i.e. their mother tongue, was B1, which is the level expected for a foreign language. This has serious implications for teaching Arabic in UAE and other Arab countries. More rigorous studies are needed to find out the average level of students' language proficiency at different levels, as it would impact their ability to understand and express certain academic concepts. This in turn could lead to reviewing Arabic learning objectives and teaching methodologies. It has been noted that students' proficiency in writing may

be affected by their exposure to the language of social and digital media. However, new technologies available to students, such as Generative Artificial Intelligence tools may have more profound impacts on how they learn the language and use it in academic and other contexts.

The Arabic learning and teaching context is very complex, and trying to understand attainment of students in different countries is a major challenge, with the absence of a framework like the CEFR. The Arab world needs a framework which helps in language policy and planning, informs decision making about curricula and learning objectives, and helps with the accreditation and mutual recognition of qualifications across the Arab region, as was the case for the CEFR and the EU. This framework needs to address the specific challenges that Arabic learners face due to Arabic diglossia, the grammatical complexity of the language, and other features that distinguish Arabic from European languages for which the CEFR was created, as well as the different socio-cultural contexts within the Arab region. If the CEFR is to be used as the basis of this endeavour, radical adaptation will be necessary for Arabic; alternatively, a similar framework could be developed specifically. Nonetheless, the need for such framework exists, and it is urgent. The existence of an Arabic translation of the CEFR is welcomed as a positive step. However, it is also clear that a tailored framework that considers the specific features of the Arabic language needs to be developed.

7 References

Alderson, J. Charles. 2007. The CEFR and the need for more research. *Modern Language Journal* 91(4). 695-663.

Al-Jarf, Reima & Nilya G Mingazova. 2020. Evaluation of Arabic language teaching textbooks used in Russia in the light of the CEFR criteria. Proceedings of the International Forum on Teacher Education IFTE-2020, 101-129. https://doi.org/10.3897/ap.2.e0101.

Badawi, El-Said M. 1973. *Mustawayaat ʻal ᶜarabiyya ʻal-muᶜaasira fi Misr* [Levels of Contemporary Arabic in Egypt]. Cairo: Dar ʻal-Maᶜaarif.

Bies, Ann, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander, Owen Rambow. 2014. *Transliteration of Arabizi into Arabic orthography: Developing a parallel annotated Arabizi-Arabic script SMS/chat corpus*. In: Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar. Association for Computational Linguistics, 93–103. https://www.aclweb.org/anthology/W14-3612.pdf (accessed 27July 2023).

Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Strasbourg: Council of Europe. www.coe.int/lang-cefr (accessed 27July 2023).

Ferguson, Charles A. 1959. Diglossia. *Word* 15. 325-340.

Foley, Joseph A. 2019. Issues on Assessment using CEFR in the Region, *LEARN Journal: Language Education and Acquisition Research Network Journal*, 12(2), 28-48. https://so04.tci-thaijo.org/index.php/LEARN/article/view/195583 (accessed 25 July 2023).

Habash, Nizar and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 79–88, Marseille, France: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.9.pdf (accessed 23 July 2023).

Hazar, Esin. 2021 The Influence of the CEFR in Turkish National Curriculum, *African Educational Research Journal* 9(2). 551-561.

Horn, Christopher. 2015. Diglossia in the Arab World. *Open Journal of Modern Linguistics* 5. 100-104. http://dx.doi.org/10.4236/ojml.2015.51009.

Hulstijn, Janh. 2007. The shaky ground beneath the CEFR quantitative and qualitative dimensions of language proficiency. *Modern Language Journal* 91(4), 663-667

Kaye, Alan S. 2001. Diglossia: The state of the art. *International Journal of the Sociology of Language* 152. 117-129.

Lu, Yang. (Ed.). 2017. *Teaching and Learning Chinese in Higher Education: Theoretical and Practical Issues* (1st ed.). London: Routledge. https://doi.org/10.4324/9781315520810.

Maamouri, Mohamed. (1998). Language education and human development: Arabic diglossia and its impact on the quality of education in the Arab region. The Mediterranean Development Forum. Washington, DC: The World Bank

Mallek, Fatma, Billal Belainine & Fatiha Sadat. 2017 Arabic social media analysis and translation. *Procedia Computer Science* 117. 298–303. https://reader.elsevier.com/reader/sd/pii/S1877050917321786?token=04D78450201150EB2BE784E24FF9369AC7E2C453677C498269A57445799E2E405D67910B49D22661BEA40C9C6C153759 (accessed 27 July 23).

Marçais, William. 1930. La diglossie arabe [Arabic Diglossia]. *L'Enseignement Public*. 97. 401-409.

Mohamed, Salwa. 2021. An Analysis and Levelling of Reading Texts Across Arabic Textbooks Based on the CEFR Proficiency Levels. In Xiang, C.H. (ed.), *Trends and developments for the future of language education in Higher Education,* 57-83. Hershey, Pennsylvania): IGI Global.

Mohamed Salwa. 2023. The development of an Arabic curriculum framework based on a compilation of salient features from CEFR level descriptors. *The Language LearningJournal* 51(1). 33-47. https://doi.org/10.1080/09571736.2021.1923781.

Neff-van Aertselaer, J.A. 2013, Contextualizing EFL argumentation writing practices within the Common European Framework descriptors. *Journal of Second Language Writing* 22. 198–209.

Negishi, M. 2012. The Development of the CEFR-J: Where we are, where we are going. In N. Tomimori, M. Furihata, K. Haida, N. Kurosawa, & M. Negishi (Eds.), *New perspectives for foreign language teaching in higher education: Exploring the possibilities of application of CEFR*,105-116. Tokyo: Tokyo University of Foreign Studies. http://www.tufs.ac.jp/common/fs/ilr/EU_kaken/_userdata/negishi2.pdf (accessed 23 June 23).

North, Brian. 2022. Case studies in mediation, online interaction and plurilingualism: Implementing the CEFR Companion Volume. *TEANGA, Journal of the Irish Association for Applied Linguistics* 29. 1–12. https://doi.org/10.35903/teanga.v29i.3774.

North, Brian, Enrica Piccardo, Tim Goodier, Daniela Fasoglio, Rosanna Margonis-Pasinetti & Bernd Rüschof. 2022. *Enriching 21st Century Language Education: The CEFR Companion Volume in Practice*. Strasbourg: Council of Europe.

Piccardo, Enrica & Brian North. 2022. Enriching the Scope of Language Education: The CEFR – Companion Volume. In North, Brian, Enrica Piccardo, Tim Goodier, Daniela Fasoglio, Rosanna Margonis-Pasinetti & Bernd Rüschoff. (eds.), *Enriching 21st Century Language Education: The CEFR Companion Volume in Practice*. Strasbourg: Council of Europe. 27-48

O'Dwyer, Fergus, Morten Hunke, Alexander Imig, Noriko Nagai, Naoyuki Naganuma and Maria Gabriela Schmidt. 2017. *Critical, Constructive Assessment of CEFR-informed Language Teaching in Japan and Beyond*. Cambridge: Cambridge University Press.

Ryding, Karin C. 1991. Proficiency despite diglossia: A new approach for Arabic. *Modern Language Journal* 75. 212-218.

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Soliman, Rasha. 2018. The implementation of the Common European Framework of Reference for the teaching and learning of Arabic as a second language in higher education. In: Wahba, Kassem M., Liz England, Zeinab A. Taha. (eds.), *Handbook for Arabic Language Teaching Professionals in the 21st Century, Volume II,* 118-137. New York: Routledge.

Versteegh, Kees. 2001. *The Arabic Language*. Edinburgh: Edinburgh University Press.

8 Biography

Dr. Aziza Zaher is Associate Professor in Arabic Studies at Durham University's School of Modern Languages and Cultures and a Senior Fellow of the Higher Education Academy. She holds a PhD in critical discourse analysis, MA in linguistics, and a Postgraduate Certificate in Learning and Teaching in Higher Education. Her research interests include teaching Arabic as a foreign language, translation, linguistics, discourse analysis, critical discourse analysis, and the application of CEFR to Arabic. She has co-authored several publications, including three books: *Working with Arabic Prepositions* (2020), *Chambers Arabic Vocabulary* (2009), and *Chambers Arabic Grammar* (2009), and published several articles and book chapters. She has delivered keynotes, organised workshops and given presentations at many conferences.

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Responding to the CEFR Alignment Handbook: Sharing experience of alignment activities and reflecting on lessons learned

The *Handbook* Steering Group:

Neus Figueras, European Association for Language Testing and Assessment (EALTA)

David Little, Trinity College Dublin

Barry O'Sullivan, British Council

Nick Saville, Association of Language Testers in Europe (ALTE)

Lynda Taylor, UK Association for Language Testing and Assessment (UKALTA)

https://doi.org/10.37546/JALTSIG.CEFR7-6

In February 2020, a conference entitled "The CEFR: Towards a Road Map for Future Research and Development", co-sponsored by EALTA, UKALTA and the British Council, was hosted by the British Council in London (O'Dwyer et al. 2020, Little and Figueras 2022). The organizing institutions recognized the need to explore ways of developing research methodologies and projects of various kinds that could help to extend and further develop the CEFR and its implementation. It was hoped that the conference would inform the development of a road map for future engagement with the CEFR, taking account of what had been learnt so far and of new developments in applied linguistics and related disciplines.

Participants in the February 2020 conference agreed on the need for a new *Handbook* to support the alignment of language education with the CEFR and its Companion Volume. Accordingly, the three organizations behind the conference, together with ALTE, developed *Aligning Language Education with the CEFR: A Handbook*, publishing it online in April 2022 (Figueras et al. 2022). In undertaking to produce the *Handbook*, the steering group (which emerged from the February 2020 event) recognized that alignment applies not only to language tests but to policy, curriculum guidelines, curricula, syllabuses, textbooks and other teaching/learning resources. The group also decided that the *Handbook* should serve to inform policy makers, teacher educators, teachers and other language education stakeholders, as well as supporting the more or less technical processes on which alignment depends. The *Handbook* also seeks to help users to navigate the wide range of CEFR-related reference documents.

In the expectation that the *Handbook* would generate new interest in CEFR alignment practices, the

editors stated in their foreword that they planned to organize a conference in 2024 to give language education professionals an opportunity to share their alignment projects and reflect on the usefulness of the *Handbook*. The conference was held on 18 and 19 October 2024 at Blanquerna – Universitat Ramon Llull, Barcelona. It was hosted by Cristina Corcoll and her colleagues in the GREDA research group at Blanquerna, and organized by the *Handbook* steering group (Neus Figueras, David Little, Barry O'Sullivan, Nick Saville, Lynda Taylor). Notwithstanding the travel restrictions imposed by many academic institutions and the political situation in Europe and beyond, the conference attracted more than 120 participants from 24 countries and four continents. The programme comprised five plenary sessions, 35 presentations and eight posters. The presentations included case studies from China, Costa Rica, Cuba, Ireland, Japan, Ukraine, Saudi Arabia, United Kingdom and Italy. Presenters' PowerPoint slides are available at https://ealta.eu/wp-content/uploads/2024/12/CEFR_Handbook_programme_overview_final.pdf.

Sarah Breslin, Executive Director of the European Centre for Modern Languages (ECML) and Head of Language Policy at the Council of Europe in Strasbourg, thanked the organizers for inviting her to attend the conference and learn more about critical engagement with the CEFR in general and CEFR alignment projects in particular; such events are a welcome response to the Council of Europe's efforts to support innovation in language education.

In the opening plenary, *Perspectives on CEFR alignment*, David Little (Trinity College Dublin) and Constant Leung (King's College London) set the scene in a conversation that addressed the CEFR's foundations and the challenges that the CEFR and CEFR CV present. David Little drew attention to the CEFR's double inheritance – the language user/learner as autonomous social agent and the scaled description of L2 proficiency designed to facilitate cooperation among Council of Europe member states. He argued that this double inheritance implies two possible approaches to alignment, one that starts from the learner as plurilingual social agent and one that starts from communicative language activities and competences, levels of proficiency and scales. Constant Leung focused on relevant research in related fields – community interpreting, medical and professional communication, interactional competence, scenario-based assessment – which suggests that there is a good educational and intellectual case for the cohabitation of the two orientations identified by David Little.

In the second plenary, *Views on CEFR alignment from the publishing perspective*, Ben Knight from Oxford University Press and David Bradshaw from Cambridge University Press and Assessment responded to questions raised by the session chair, Lynda Taylor, on the relevance of the CEFR for them as publishers and on how they address alignment in their context. For publishers, the most important sources of information when developing materials for a specific country are ministry guidelines or a national curriculum, which may not be aligned with the CEFR. The 'takeaway message' from both speakers was that ELT publishers take the alignment of their language learning materials with the CEFR very seriously, although CEFR alignment happens behind the scenes, using curriculum frameworks and editorial guidelines. Most customers prefer to use their own judgement on the appropriacy of those materials for their students rather than to read documentation about the alignment process.

The plenary at the end of the first day, *Ideas shared and lessons learnt on Day 1*, presented participants with the following list of topics covered and issues raised in the different sessions:

- The history of the CEFR, its legacy and its evolution
- The emerging/evolving nature of the construct/s of language & communication
- The value of cross-disciplinary engagement
- The need for adaptation, customisation, tailoring of the CEFR
- The critical importance of 'context'
- The importance of shared understanding, language and discourse
- The CEFR - & CEFR alignment - as an instrument for social justice

The second day of the conference started with the fourth plenary, *The future of CEFR alignment supported by emerging technologies*, where Barry O'Sullivan from the British Council and Nick Saville from ALTE asked conference participants to consider what technologies are most used in alignment projects and in what ways technology and artificial intelligence can enhance alignment of language education. Discussion followed on the affordances, risks and challenges associated with using technology and AI and on whether and how the *Handbook* could take those on board.

The fifth plenary session, *Aligning tests and testing systems in context*, comprised four presentations: 'A flexible, inclusive approach to a statutory CEFR alignment requirement for EL centres in Ireland' (Elaine Boyd, Thom Kiddle and Mary Grennan, Quality and Qualifications Ireland), 'Mapping the SMEEA Gaokao tests to the CEFR: multilingual alignment using the new *Handbook*' (Graham Seed, Cambridge University Press and Assessment), 'Implementation, use and future of CEFR in some countries in Latin America' (Walter Araya, Unversidad de Costa Rica and LAALTA), and 'Reflections on the *Handbook*: three stories from Japan' (Masashi Negishi, Tokyo University of Foreign Studies).

The sixth plenary session, 'The CEFR across educational contexts and systems', comprised three presentations: 'A report from Japan' (Masashi Negishi, and Yukio Tono, Tokyo University of Foreign Studies), 'Translating the CEFR CV and the CEFR Alignment *Handbook*' (Javier Fruns, Instituto Cervantes), and 'CEFR Journal: Creating dialogue between research and practice' (Morten Hunke, Brandenburg University of Applied Sciences and Humboldt-Universität zu Berlin)

The closing session, *Reflecting on the conference and considering next steps,* highlighted the following three issues, which stimulated lively discussion:

- The role that technology might play in the near future
- How context shapes alignment processes and CEFR uses in Europe and beyond
- Dissemination initiatives reinforcing collaboration

At the end of the conference participants were asked to use Slido to say in one or two words how they were feeling. The results are shown in Figure 1.

Figure 1: How participants said they were feeling at the end of the conference

In the course of the conference, it became clear that the *Handbook* provides useful guidance but also that it needs to be more widely disseminated. The notes made by the session chairs, the reactions from participants to the presentations, and the contributions from the audience in the plenary sessions indicated that further work on the *Handbook* should include reflection on changing linguistic

and cultural topographies, additional navigational advice on how to get from A to B, and alternative alignment methods, modes and tools (online, digital, corpora, AI, …). There was general agreement on the importance of maintaining terminological coherence across languages. The *Handbook* has been translated into Spanish by the Cervantes Institute and German by the Goethe Institute; Greek and Arabic translations are in progress.

Drawing on the rich body of research shared at the conference, the steering group plans further work on the *Handbook*, setting up short, medium and long-term objectives. The short-term objectives are to develop a supplement to the 2022 Handbook, which will address issues raised at the Barcelona conference, and to guest-edit a special issue of the CEFR Journal devoted to the alignment of language education with the CEFR. These short-term initiatives will be accompanied by efforts to reach out via international networks to language teacher educators and language teachers.

References

Figueras, Neus, David Little & Barry O'Sullivan. 2022. Aligning Language Education with the CEFR: A Handbook. *CEFR Journal – Research and Practice* Volume 5: 5–10. https://doi.org/10.37546/JALTSIG.CEFR5-1

Little, David & Neus Figueras (eds.). 2022. *Reflecting on the Common European Framework of Reference for Languages and its Companion Volume*. Bristol: Multilingual Matters.

O'Dwyer, Fergus, Morten Hunke & Maria Gabriela Schmidt. 2020. The EALTA UKALTA "Roadmap" conference – The CEFR: a road map for future research and development – meeting overview. *CEFR Journal – Research and Practice* Volume 2: 89–97.

Biographies

Neus Figueras coordinated the development of foreign language curricula and certificate examinations for adult language learners in the Catalan ministry of education in Spain for over 20 years. She has been involved in a number of international research and development projects related to assessment at different education levels. She collaborates regularly with the Council of Europe in the uses and the dissemination of the CEFR. She is a founding member of EALTA (European Association for Language Testing and Assessment) and was the first president of the association. She is currently the coordinator of the EALTA CEFR SIG.

David Little is a Fellow Emeritus of Trinity College Dublin, Ireland. His principal research interests are the theory and practice of learner autonomy in second language education and the management of linguistic diversity in schools and classrooms. He has been a regular contributor to the Council of Europe's language education projects since the 1980s, especially in relation to the European Language Portfolio. In 2010, the National University of Ireland awarded him an honorary doctorate in recognition of his contribution to language education in Ireland and further afield.

Barry O'Sullivan is the Head of Assessment Research & Development at the British Council. He has undertaken research across many areas on language testing and assessment and its history and has worked on the development and refinement of the socio-cognitive model of test development and validation since 2000. He advises ministries and institutions on assessment policy and practice and is particularly interested in the communication of test validation and in test localisation. He is the founding president of the UK Association of Language Testing and Assessment (UKALTA) and holds honorary and visiting chairs at a number of universities globally.

Nick Saville is Director of Thought Leadership at Cambridge University Press & Assessment (English) and Secretary-General of the Association of Language Testers in Europe (ALTE). He graduated in Linguistics in 1980 and holds an MA TEFL and an MA (Cantab). He wrote his thesis on the impact of language assessment in educational contexts supervised by Prof. Cyril Weir at the University of Bedfordshire.

Nick taught English at the University of Cagliari (Italy) for 6 years from 1980 and began working for the University of Cambridge in 1987, where he had several roles related to test development before being appointed Director of Research and Validation in 2001. He has taken part in many projects with the European institutions - the Council of Europe, the European Parliament & the EU Commission and his areas of expertise include: English language learning and plurilingualism; Learning Oriented Assessment (LOA) and the integration of learning with assessment (ILA); Ed*Tech* combined with AI for the automation of language assessment; language assessment literacy; impact research and ethical frameworks in language education. He currently co-edits the *Studies in Language Testing* series (CUP) and in 2023, he became a Fellow of the Academy of Social Sciences (FAcSS) in UK.

Lynda Taylor is Visiting Professor at the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire, UK. She has worked for nearly 40 years in the field of language teaching, learning and assessment, first as an ELT teacher and teacher educator, later as a materials developer and assessment researcher, particularly with IELTS and the full range of Cambridge English qualifications. Her research interests include speaking and writing assessment, testing accommodations for language learners with special needs and the development of language assessment literacy. She was formerly Assistant Research Director with Cambridge Assessment English and has advised on test development and validation projects around the world, including Europe, North America and Asia. She has presented invited plenaries, research papers and workshops internationally, and has published extensively in academic journals. She authored or edited many of the volumes in CUP's *Studies in Language Testing* series. In 2022 she was awarded Fellowship of the UK Academy of Social Sciences and has been President of the UK Association for Language Testing and Assessment (UKALTA) since 2020. Her most recent publications include a pair of co-edited volumes on the topic of *Language Assessment Literacy and Competence* (CUP).

CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Title: **CEFR Journal – Research and Practice**

Type: Online journal
URL: https://cefrjapan.net/journal
Contact: journal@cefrjapan.net

Edited by: Japan Association for Language Teaching (JALT)
CEFR & Language Portfolio SIG
Fergus O'Dwyer (editor in chief)
Dmitri Leontjev (editor)
Elif Kantarcıoğlu (editor)
Maria Gabriela Schmidt (coordinator, editor)
Morten Hunke (liaison officer, editor)
Alexander Imig (treasurer, website editor)

ISSN: 2434-849X
DOI https://doi.org/10.37546/JALTSIG.CEFR6

Publication Statement of the CEFR Journal

All work published in the CEFR journal is original work, with contributions from the named authors.

The work published in the CEFR journal is open access and licensed under a Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

Authors of articles are allowed to retain publishing rights and hold the copyright without restrictions. The CEFR Journal actively checks for plagiarism using Turnitin. The CEFR Journal does not charge article processing charges or submission charges.

Any issues regarding complaints and appeals, conflicts of interest / competing interests; ethical oversight; post-publication discussions and corrections should be directed to journal@cefrjapan.net

The administration costs of the CEFR Journal are covered by the JALT CEFR & Language Portfolio SIG; the CEFR Journal has no paid advertisements or other revenue sources.

In the event that the CEFR Journal is no longer published, electronic backup and preservation of access to the journal content shall be made available via CLOCKSS.

Submission (Call for Papers)

This journal attempts to fall somewhere in between an inaccessible academic journal (long waiting times, fairly strict guidelines/criteria) and a newsletter (practical in nature but lacking in theoretical support/foundation), linking research of a practical nature with relevant research related to foreign language education, the CEFR, other language frameworks, and the European Language Portfolio. While the CEFR was introduced by the Council of Europe and intended for use, first and foremost, within Europe, the influence of the CEFR now has to be attested in many places beyond European borders. It has become a global framework, impacting a variety of aspects of language learning, teaching, and assessment across countries and continents beyond the context for which it was originally created. As such, there is a pressing need to create a quality forum for sharing research, experiences, and lessons learned from applying the CEFR in different contexts. This journal provides such a forum where people involved or interested in processes of applying the CEFR can share and learn from one another.

We are continuously seeking contributions related to foreign language education, the CEFR, other language frameworks, and the European Language Portfolio. We are particularly interested in specific contextual adaptations.

Until 30 June 2025 we are looking for abstracts at: journal@cefrjapan.net.

Please contact the editors with any queries and submit to: journal@cefrjapan.net

Guidelines

Submission:	30 June 2025
Contributions:	Articles (research), reports (best practice), news (work in progress), research notes, book reviews
Language(s):	English (British, American, international) preferred, but not mandatory. Other languages by request, with an extended abstract in English.
Review type:	Peer review, double blind

Reviewer guidelines:

We ask all reviewers to make every reasonable effort to adhere to the following ethical guidelines for **CEFR Journal – Research and Practice** articles they have agreed to review:

1. Reviewers must give an unbiased consideration to each manuscript submitted for consideration for publication, and should judge each on its merits. Since, we employ a double-blind review, the text you have been provided with ought to have no reference to race, religion, nationality, sex, gender, seniority, or institutional affiliation of the author(s). Please, notify us immediately where any such information is still detectable in the anonymised text you received.
2. Reviewers should declare any potential conflict of interest prior to agreeing to review a manuscript, including any relationship with the author that may potentially bias their review.

3. Reviewers are strongly advised to keep the peer review process confidential; information or correspondence about a manuscript should not be shared with anyone outside the peer review process.
4. Reviewers should provide a constructive, comprehensive, evidenced, and appropriately substantial peer review report. For your convenience, we are providing you with a 'reviewing matrix' you may choose to use at your own discretion. We would also like to kindly ask you to provide us in the journal editorial team with a final overall assessment of the text's publication potential – please, see bottom of this document.
5. Reviewers must avoid making statements in their report, which might be construed as impugning any person's reputation.
6. Reviewers should make all reasonable effort to submit their report and recommendation in a timely manner, informing the editor if this is not possible.
7. Reviewers should call to the journal editor's attention any significant similarity between the manuscript under consideration and any published paper or submitted manuscripts of which they are aware.

Author instructions:

Adapted version of deGruyter Mouton guidelines for Language Learning in Higher Education (CercleS) and style sheet.

CEFR Journal—Research and Practice

Japan Association for Language Teaching (JALT)
CEFR & Language Portfolio SIG (CEFR & LP SIG)

ISSN 2434-849X