RESEARCH AND ANALYSIS

# Investigating standards in GCSE French, German and Spanish through the lens of the CEFR

Milja Curcin and Beth Black

ofqual

# Acknowledgements

We would like to thank the many people without whose work or advice this study would not have been possible:

- all our participants, who devoted a lot of their time and enthusiasm to work on this study and share their expertise and opinions,
- colleagues at Ofqual who have helped in different ways (with IT support, admin and paper shuffling, analytical support and advice, and various ad hoc and last minute request for help) – in particular, Nadir Zanini, Joe Colombi, Robin Smith, Ben Laurens, Matthew Stratford, Richard Coles and Jonathan Clewes,
- Jane Lloyd, Alastair Pollitt, Stuart Shaw and Neil Jones, for invaluable insights and advice.

# Contents

# List of tables

# List of figures

# Executive summary

While most stakeholders would agree that modern foreign language (MFL) study is a valuable part of the curriculum, there is general decline in numbers of students taking GCSEs in these subjects. There is a persistent perception that MFL GCSEs are more difficult compared to other subjects. This is often cited as a reason for declining subject take-up at secondary and university level. On the face of it, consistent patterns in statistical evidence appear to support the notion that MFL GCSEs are graded more severely than other GCSE subjects. However, while such statistical analyses may indicate on average lower grade outcomes when controlling for prior or concurrent attainment, these analyses do not take into account a multitude of factors related to (perceptions of) difficulty and demand. These could be, for instance, subject demand, nature of assessment, allocation of teaching time and other resources, motivation of students, efficiency and effectiveness of teaching and learning, etc. (Coe, 2008; Newton, 2012; Lockyer and Newton, 2015; Wingate, 2018; Macaro, 2008; Graham, 2002; Klapper, 2003; etc.).

This study was part of a programme of research carried out by Ofqual to help inform its policy decision of whether to intervene and adjust grading standards in MFL GCSE qualifications in French, German and Spanish. The study was designed to describe the nature of performance and assessment standards in these subjects using the 'metalanguage' of the Common European Framework of Reference for languages (CEFR), an internationally widely used framework describing language ability via a common 'can do' scale, allowing broad comparisons across languages and qualifications. The aim was to provide a platform for a more principled discussion about whether GCSE MFL performance standards and corresponding grading standards are appropriate for these qualifications, or are indeed too high.

We do not believe that possible discrepancies between the notions of communicative language competence and language use as described in the CEFR, and the way communicative language competence and use may be understood, taught, and assessed at GCSE level, would in itself invalidate an attempt to describe GCSE MFLs in terms of CEFR descriptors. We would argue that, as long as the broad intention of the MFL GCSE curriculum and pedagogy is reasonably aligned to the CEFR – and this would appear to be the case as, for instance, MFL GCSEs should "develop [learners'] ability to communicate confidently and coherently with native speakers in speech and writing, conveying what they want to say with increasing accuracy" (DFE, 2015: 3) – a description in terms of the CEFR may not only be appropriate, but also helpful.

However, we do believe that it is important to be aware of the specific context of the MFL GCSEs, as it may account for occasional disjoint between CEFR descriptors and GCSE assessments/performances that are observed in the linking. In addition, an awareness of these discrepancies could be helpful for improving both current language pedagogy and assessment methods where appropriate, helping learners to achieve the goal of communicative language competence at the level appropriate for the phase of education at which they are.

Because this study was designed as a piece of research to answer a specific research question, rather than as a full-blown linking study, it consequently has some potential limitations in scope and generalisability. This is, to our knowledge, the first explicit attempt to link GCSE MFL qualifications to the CEFR using

recommended methodology, and so we consider this study primarily exploratory. Involvement and endorsement of other relevant stakeholders (e.g. Department for Education, exam boards), greater resources, further refinement of some aspects of the methodology and linking of specifications from other exam boards would be necessary to conduct a linking study where the results might be considered to represent an “official” linking. Therefore, the findings need to be treated as essentially descriptive and indicative. Having said this, we have made every effort to conduct this linking study according to best practice in the field, and in this sense, the results should be reasonably robust for those specifications on which the linking was performed.

In this study, key grades (grades 9, 7 and 4) in GCSE French, German and Spanish on the summer 2018 tests were notionally linked to the CEFR scale. Initially, content mapping (i.e., relating the construct and content coverage of the GCSE to the CEFR) was carried out for each subject by a CEFR expert and a GCSE subject expert. Subsequently, panels of 13 experts (including CEFR experts, Higher Education and subject experts, A level teachers and exam board representatives) carried out the following activities for each subject:

- For writing and speaking, they rank ordered, in terms of overall quality, series of GCSE performances (at grades 9, 7 and 4) interspersed with performances previously independently benchmarked on the CEFR scale. This created an overall performance quality scale on which the relative position of the GCSE and CEFR performances was determined, and CEFR-related performance standards at grades 9, 7 and 4 extrapolated from this.
- For reading and listening comprehension, they conducted a ‘standard linking’ exercise using the ‘Basket Method’ to rate each mark point on the tests in terms of the CEFR levels. CEFR level cut scores were derived from these ratings and grades 9, 7 and 4 related to these in terms of proportions of marks on the test needed to achieve each.
- The linking results at component level were averaged to get a qualification-level estimate of the mapping of each grade to the CEFR level.

The results of the linking at component level are shown in Tables 1 to 3 . The linking of GCSE grades to the CEFR levels across components within Spanish and German is very consistent, with productive skills being at a lower CEFR level than the receptive skills. French mapping is less consistent, but this may be partly due to the issues with the CEFR exemplars for productive skills, and apparent issues with the listening comprehension paper (described in the Results section). Therefore, we would suggest that the linking for French is more tentative than for the other two languages. The patterns are broadly consistent across the 3 languages, with the notable exception of grade 7 for productive skills (lowest standard in Spanish), and grade 4 for receptive skills (highest standard in Spanish).

Figure 1 shows indicative linking at qualification level for each grade, based on averaging across the CEFR sub-levels of components. It appears that performance standards between the 3 languages are reasonably aligned at qualification level despite some component-level inconsistencies. The results suggest that grade 4 is around high A1 level for Spanish and mid A1 level for German and French. Grade 7 is around mid A2 level and grade 9 around low B1 for all languages. This result accords with the results of the content mapping, which suggested that each of the 3 GCSE MFL specifications assessed most of the skills up to A2+ (i.e. high A2) level, with some aspects of language competence assessed up to low B1 level. While a degree of consistency across languages is perhaps to be expected given that these assessments are supposed to be developed based on specifications that should be reasonably aligned in terms of content and implicit demand, there is no particular reason why we should expect the performance standards for different grades to be perfectly aligned across languages. This reminds us that considering standards between even quite related subjects involves considerable nuance and interpretation.

However, in addition to the limitations discussed in the Limitations section, an important “health warning” regarding the interpretation of this linking is in order. It should be borne in mind that the limitations of assessments highlighted in both content mapping and in discussion with panellists, particularly with respect to assessment of interaction and integrated skills, would to some extent limit the interpretation based on these assessments that candidates are fully at A2 or B1 level. This is because the assessments themselves provide little evidence of some of the skills essential for communicative language competence, such as ability to engage in meaningful interaction. In a sense, it may be more appropriate to say that, overall, candidates achieving each of the GCSE grades possess most, but not all, of the skills and knowledge required of the CEFR level assigned in this linking exercise. While this is also true of A2 level to some extent, most of the caveats and discrepancies relate to where assessments appear to be targeting B1 level, as in many cases assessments were patchy in the extent to which they allowed for all of the skills relevant for B1 level to be demonstrated. This would mean that the levels assigned to different grades could be seen as overestimates to some extent, particularly for B1 level, but also to some extent for A2. This should be borne in mind in any discussions about whether A2 or B1 level may be appropriate for different GCSE grades.

This linking study dealt with describing the content/construct of GCSE MFL specifications and tests, as well as performances, in terms of the CEFR, and relating the current GCSE grading standards to the CEFR. The results essentially give an indication of where GCSE assessments are pitched and which performance standards are represented by different GCSE grades, using the language of the CEFR descriptors. Therefore, this linking is not a statement of what the GCSE standard should be, but an approximate description of what the performance and assessment/grading standard currently appears to be, using the language and descriptors of the CEFR.

The GCSE MFL assessments reviewed in this study do not appear to elicit sufficient evidence of certain linguistic skills that may be considered by some to be a crucial part of communicative language competence. It would seem important to investigate these issues further and explore ways in which the assessments might be made more effective in assessing these important skills. As far as GCSE MFLs should enable learners to act in real-life situations, expressing themselves and accomplishing tasks of different natures, it would make sense that, like the CEFR, they put the co-construction of meaning (through interaction) at the centre of the learning and assessment process.

The results are offered to stakeholders for consideration as to whether the content and performance standards and assessment demands associated with the key GCSE grades are appropriate given the purpose of GCSE qualifications, the spirit and nature of the curriculum, and the current context of GCSE MFL learning and teaching. For instance, if the relevant stakeholders were to conclude that, generally speaking, a mid A2 level of performance is too high for GCSE grade 7, this could provide rationale to support a change to grading standards. However, in this case, this rationale would not be based on statistical evidence or any notions of comparable 'value-added' between different subjects, but based on an understanding of what an appropriate performance standard, in terms of what students can do, is or should be for each grade within MFLs themselves.

We would suggest, however, in the spirit of the CEFR, that discussions around the appropriateness of language performance and assessment standards should consider important aspects of the context of language teaching in schools. The CEFR (Council of Europe, 2018: 28) suggests planning backwards from learners' real life communicative needs, with consequent alignment between curriculum, teaching and assessment. As North (2007a) points out, educational standards must always take account of the needs and abilities of the learners in the context concerned. Norms of performance need to be definitions of performance that can realistically be expected, rather than relating standards to "some neat and tidy intuitive ideal" (Clark 1987: 46). This posits an empirical basis to the definition of standards. If used appropriately, the CEFR could aid this endeavour in the context of GCSE MFLs in England.

Table 1 *GCSE to CEFR mapping for Spanish*

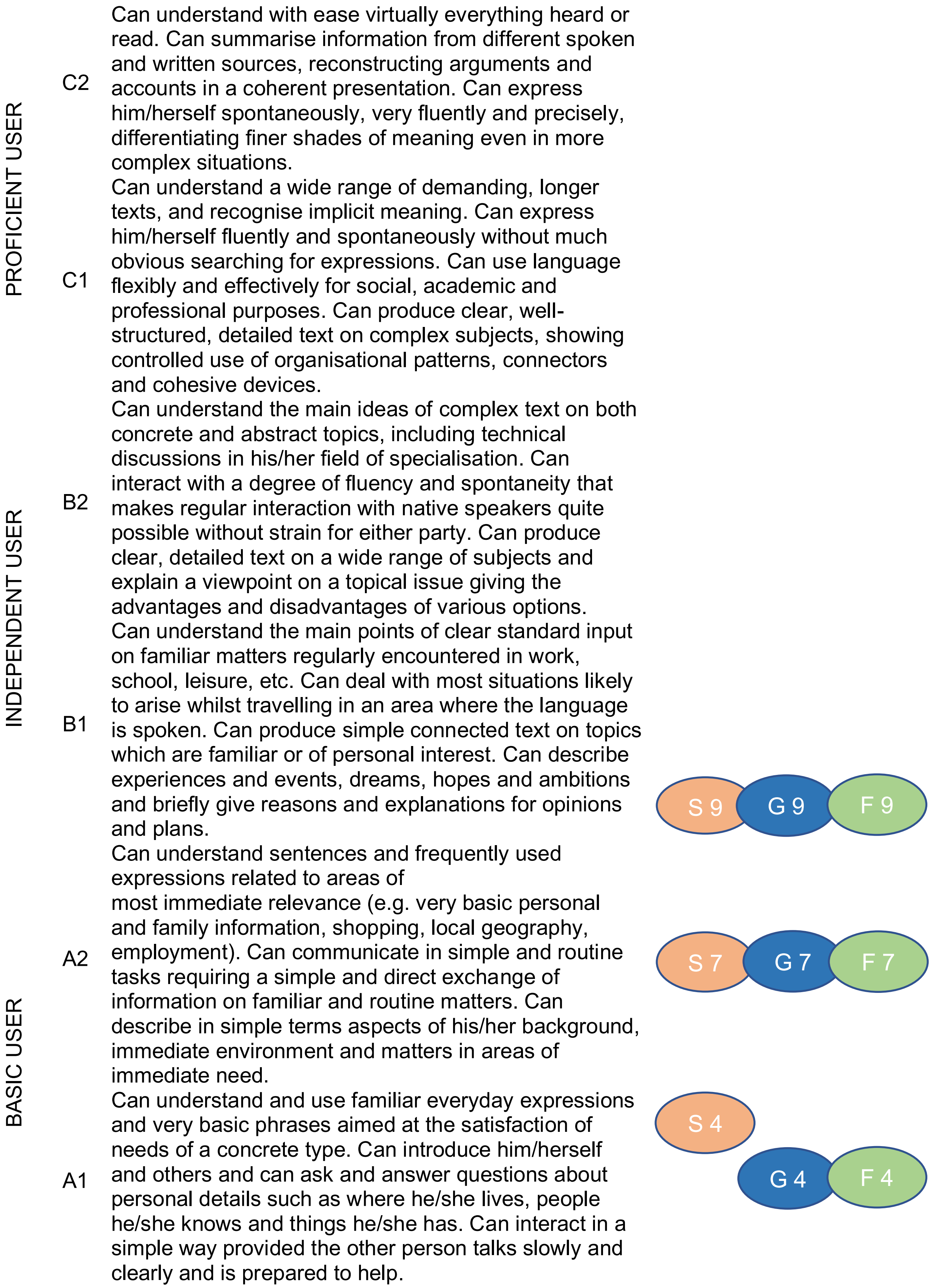
| | Writing | | Speaking | | Reading | | Listening | |
|---|---|---|---|---|---|---|---|---|
| GCSE grade | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level |
| 4 | Mid-high A1 | A1 | Low-mid A1 | A1 | Low-mid A2 | A2 | Low-mid A2 | A2 |
| 7 | Low-mid A2 | A2 | Low-mid A2 | A2 | Mid-high A2 | A2 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 | Low-mid B1 | B1 | Low-mid B1 | B1 | Low-mid B1 | B1 |

Table 2 *GCSE to CEFR mapping for German*

| | Writing | | Speaking | | Reading | | Listening | |
|---|---|---|---|---|---|---|---|---|
| GCSE grade | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level |
| 4 | Low-mid A1 | A1 | Mid A1 | A1 | High A1-low A2 | A1/A2 | High A1-low A2 | A1/A2 |
| 7 | Mid-high A2 | A2 | High A2 | A2 | Mid-high A2 | A2 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 | Low B1 | B1 | Low-mid B1 | B1 | Low-mid B1 | B1 |

Table 3 *GCSE to CEFR mapping for French*

| | Writing | | Speaking | | Reading | | Listening | |
|---|---|---|---|---|---|---|---|---|
| GCSE grade | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level |
| 4 | High A1-Low A2 | A1/2 | Low-mid A1 | A1 | High A1-low A2 | A1/A2 | Low-mid A1 | A1 |
| 7 | Low-mid B1 | B1 | High A2-low B1 | A2/B1 | Mid-high A2 | A2 | High A1-low A2 | A1/A2 |
| 9 | Low-mid B1 | B1 | Mid-high B1 | B1 | Low-mid B1 | B1 | High A2-lowB1 | A2/B1 |



| | | | |
|---|---|---|---|
| PROFICIENT USER | C2 | Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations. | |
| | C1 | Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices. | |
| INDEPENDENT USER | B2 | Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options. | |
| | B1 | Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans. | S 9 G 9 F 9 |
| BASIC USER | A2 | Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need. | S 7 G 7 F 7 |
| | A1 | Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help. | S 4<br>G 4 F 4 |

Figure 1 *Estimated qualification level mapping for each language and grade*

# Introduction

While most stakeholders would agree that modern foreign language (MFL) study is a valuable part of the curriculum, there is general decline in numbers of students taking GCSEs in these subjects. There is a persistent perception that MFL GCSEs are more difficult compared to other subjects. This is often cited as a reason for declining subject take-up at secondary and university level.

On the face of it, consistent patterns in statistical evidence appear to support the notion that MFL GCSEs are graded more severely than other GCSE subjects. However, while statistical analyses may indicate on average lower grade outcomes when controlling for prior or concurrent attainment, these analyses do not take into account a multitude of factors related to (perceptions of) difficulty and demand. These could be, for instance, subject demand, nature of assessment, allocation of teaching time and other resources, motivation of students, efficiency and effectiveness of teaching and learning, etc. (Coe, 2008; Newton, 2012; Lockyer and Newton, 2015; Cuff, 2017; Wingate, 2018; Macaro, 2008; Graham, 2002; Klapper, 2003; etc.).

This study was part of a programme of research carried out by Ofqual to help inform its policy decision of whether to intervene and adjust grading standards in MFL GCSE qualifications in French, German and Spanish. The study was designed to describe the nature of performance and assessment standards in these subjects using the 'metalanguage' of the Common European Framework of Reference for languages (CEFR), an internationally widely used framework describing language ability via a common 'can do' scale, allowing broad comparisons across languages and qualifications. The aim was to provide a platform for a more principled discussion about whether GCSE MFL performance standards and corresponding grading standards are appropriate for these qualifications, or are indeed too high.

## Why look at GCSE performance and assessment standards in relation to grading severity using CEFR descriptors

The assessment instruments and test specifications interpret GCSE MFL standards in a particular way, by including certain curriculum domains, assessment methods, marking criteria and questions of varying types and demand, guided by Department for Education subject content (DfE, 2015) and guidelines about desirable features of assessments. However, in the absence of clear and sufficiently detailed performance descriptors for different grades, it is difficult to establish whether these assessments are appropriately 'pitched' to test at appropriate and agreed level.[1]

Currently, as in other GCSEs, the grading standard of GCSE MFLs is maintained using the comparable outcomes approach, which maintains the 'value-added'

[1] Before the reformed GCSEs were sat for the first time, Ofqual, working with subject experts and senior examiners from exam boards, developed grade descriptions for grade 8, 5 and 2. https://www.gov.uk/government/publications/grade-descriptors-for-gcses-graded-9-to-1
The aim of these grade descriptions was to give teachers an indication of the likely level of performance. They were not intended to be used to set standards in the first new awards, and the intention was to review them once the new qualifications had settled down.

relationship for the cohort between Key stage 2 and GCSE. However, there is little clarity as to what it is that students at different GCSE grades should be able to do, or can actually do with language. It is also difficult to say whether GCSE assessments themselves are pitched at an appropriate level of demand, as it is not universally understood or accepted amongst stakeholders what is actually an appropriate or realistic level of demand for this qualification and individual grades. Furthermore, there is a lack of clarity with respect to what different stakeholders might consider to be appropriate requirements and performance standards for different GCSE grades (cf. the results of stakeholder surveys presented in Curcin and Black, 2019). Part of this lack of clarity is probably due to the difficulties associated with articulating performance standards in the first place.

This is primarily what our study tried to establish – where GCSE assessments are pitched and what performance standards are represented by different GCSE grades. A very useful and well-established tool for articulating performance standards in languages is the CEFR. This framework is intended to provide a ‘universal’ metalanguage for description of language competence. Once we understand which performance standards that are expected at different grades, we can then discuss whether that level is appropriate for the current context of GCSE MFL learning and teaching, given the spirit and nature of the curriculum, for different purposes of GCSEs, for different stakeholders, etc.

We are conscious that, while the CEFR is intended to provide a metalanguage for description of language competence, it is not intended to be used indiscriminately and without regard to local context and local educational aims (see below for more details on this). In this study, we took care to acknowledge the limitations of the CEFR application to the context of GSCE assessments, for instance where GCSE underspecifies certain aspects of linguistic competence at some CEFR levels, while fully according with other aspects. These will be clearly pointed out and relevant caveats highlighted in reporting the results of our linking and in any further discussions regarding the appropriate performance standards for GCSE MFLs.

It is important to emphasise that this study dealt with describing the content/construct of GCSE MFL specifications and tests, as well as performances, in terms of the CEFR, and relating current GCSE grading standards to the CEFR. This study is not a statement of what the standard should be, but an approximate description of what the performance and assessment/grading standard currently appears to be, using the language and descriptors of the CEFR.

Furthermore, we should emphasise that the GCSE to CEFR ‘linking’ attempted in this study can be considered exploratory and preliminary, rather than as an ‘official’ linking, being limited in scope to a subset of the relevant specifications. This was a research exercise, carried out to facilitate a resolution of the debates around grading severity, rather than with official linking as its main goal. Methodological and other limitations are discussed at some length in the Limitations section and in the Discussion.

## Why CEFR can be considered appropriate for use in the context of GCSE MFLs in England

The CEFR aims to describe what students can do with language (any [European] language, not just English) at different competence levels, across 6 levels of proficiency spanning from A1 (Basic User – 'Breakthrough') to C2 (Proficient User – 'Mastery') – see Figure 2. CEFR descriptors were initially developed in a multi-lingual environment, and in relation to 3 foreign languages (English, French, German) (North, 1998, 2007a, 2007b)[2] rather than solely with reference to English as the second language. Furthermore, they assume the cognitive and social competences of young adults at age 16 and above, and are thus age-appropriate for use in the context of GCSEs.

| | | |
|---|---|---|
| PROFICIENT USER | C2 | Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations. |
| | C1 | Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices. |
| INDEPENDENT USER | B2 | Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options. |
| | B1 | Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans. |
| BASIC USER | A2 | Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need. |
| | A1 | Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help. |

Figure 2 *The CEFR global scale*

According to the CEFR document (Council of Europe, 2001), the CEFR does not inherently impose any standards on the local context. It is a descriptive tool and is intended to provide a shared basis for reflection and communication among those involved in teacher education and in the elaboration of language syllabuses,

[2] See Appendix A for a brief summary of the history of the CEFR and the development of its descriptor scales.

curriculum guidelines, textbooks, examinations, etc., across different countries and educational systems. It should allow users to reflect on their decisions and practice, and to situate and co-ordinate their efforts, as appropriate, for the benefit of language learners in their specific contexts. It is a flexible tool to be adapted to the specific context of use.

North (2007a) points out that there is no need for a conflict between using a common framework such as the CEFR to provide transparency and coherence and the need to have local strategies that provide learning goals specific to particular contexts. The main danger is a simplistic interpretation of the common framework. The key to its valid use is for users to appreciate that a common framework is a descriptive metasystem that is intended as a reference point, not as a tool to be implemented without further elaboration and adaptation to local circumstances (see also, e.g. Taylor, 2004). According to North (2007b), the idea is for users to divide or merge activities, competences, and proficiency stepping-stones, as described in the CEFR, that are appropriate to their local context. The use of CEFR descriptors allows these to be related to the greater scheme of things and thus communicated more easily to colleagues in other educational institutions and, in simplified form, to other stakeholders.

Since its launch in 2001, the CEFR has been translated into approximately 30 languages. It has become the most commonly referenced document upon which language teaching and assessment has come to be based, both in Europe and internationally (O'Sullivan, 2015). An example of its international use is in Taiwan (Wu & Wu, 2010, p. 205), where all nationally recognised examinations must demonstrate a link to the CEFR. Other examples of linking for a range of different languages and tests include: Dutch foreign language state examinations (French, German and English as foreign languages); Asset languages in England; Certificate of Italian as a Foreign language; European Consortium for the Certificate of Attainment in Modern Languages (ECL) tests of German, English and Hungarian as foreign languages; Test of German as a Foreign Language (TestDaF); the City & Guilds Communicator examination, etc. (all presented in Martyniuk, W. (ed.), 2010). Furthermore, UK Quality Code for Higher Education (2015: 7) acknowledges that the CEFR has become the predominant international standard, and the Subject Benchmark Statement in this document attempts to adopt the CEFR as appropriate to UK higher education, advocating its use as a benchmark for standards of achievement at different levels in university language learning programmes (ibid.: 22). A number of university MFL departments and university language centres in England have either explicitly mapped their courses to the CEFR or make reference to the CEFR in describing the achievement levels of their students at the end of their courses.[3]

[3] https://www.nottingham.ac.uk/clas/documents/language-achievement-levels.pdf
https://www.lancaster.ac.uk/study/undergraduate/courses/modern-languages-ba-hons-r800/#structure
https://www.city.ac.uk/study/courses/short-courses/modern-languages
https://warwick.ac.uk/fac/arts/modernlanguages/intranet/undergraduate/courseoutlines/r9q1/faq/
http://www.open.ac.uk/courses/qualifications/q30
https://www.york.ac.uk/lfa/courses/long/
http://www.bristol.ac.uk/sml/study/uwlp/
https://www.kcl.ac.uk/study/undergraduate/courses/german-with-a-year-abroad-ba
https://www.brookes.ac.uk/courses/undergraduate/applied-languages
https://www.langcen.cam.ac.uk/culp/culp-general-courses.html

According to the CEFR document (Council of Europe, 2001), the CEFR comprehensively describes what language learners have to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively in that language. The description also covers the cultural context in which language is set. The CEFR also defines levels of proficiency which allow learners' progress to be measured at each stage of learning. According to the CEFR, any form of language use and learning could be described as follows (Council of Europe, 2001: 9):

> *Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of* ***competences****, both* ***general*** *and in particular* ***communicative language competences****. They draw on the competences at their disposal in various contexts under various* ***conditions*** *and under various* ***constraints*** *to engage in* ***language activities*** *involving* ***language processes*** *to produce and/or receive* ***texts*** *in relation to* ***themes*** *in specific* ***domains****, activating those* ***strategies*** *which seem most appropriate for carrying out the* ***tasks*** *to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences.*

Communicative language competence can be considered as comprising 3 key components: linguistic, sociolinguistic and pragmatic. Each of these components is postulated as comprising knowledge and skills and know-how. The language learner/user's communicative language competence is activated in the performance of the various language activities, involving reception, production,[4] interaction[5] or mediation[6] (in particular interpreting or translating). Each of these types of activity is possible in relation to texts in oral or written form, or both. This is summarised in Figure 3:

[4] The skills of writing and speaking are usually referred to as productive skills or production. The skills of listening and reading comprehension are usually referred to as receptive skills or reception.

[5] According to Council of Europe (2018: 81), Interaction, which involves 2 or more parties co-constructing discourse, is central in the CEFR scheme of language use. Spoken interaction is considered to be the origin of language, with interpersonal, collaborative and transactional functions. Interaction is also seen as fundamental in learning. The CEFR scales for interaction strategies reflect this with scales for turn-taking, cooperating (collaborative strategies) and asking for clarification.

[6] According to the CEFR text (ibid.: 14), written or oral mediation makes communication possible between persons who are unable to communicate with each other directly. Translation or interpretation, a paraphrase, summary or record, provides for a third party a (re)formulation of a source text to which this third party does not have direct access. The Council of Europe (2018: 103), expands on this definition to state that in mediation, the user/learner acts as a social agent who creates bridges and helps to construct or convey meaning, sometimes within the same language, sometimes from one language to another (cross-linguistic mediation). The focus is on the role of language in processes like creating the space and conditions for communicating and/or learning, collaborating to construct new meaning, encouraging others to construct or understand new meaning, and passing on new information in an appropriate form. The context can be social, pedagogic, cultural, linguistic or professional.

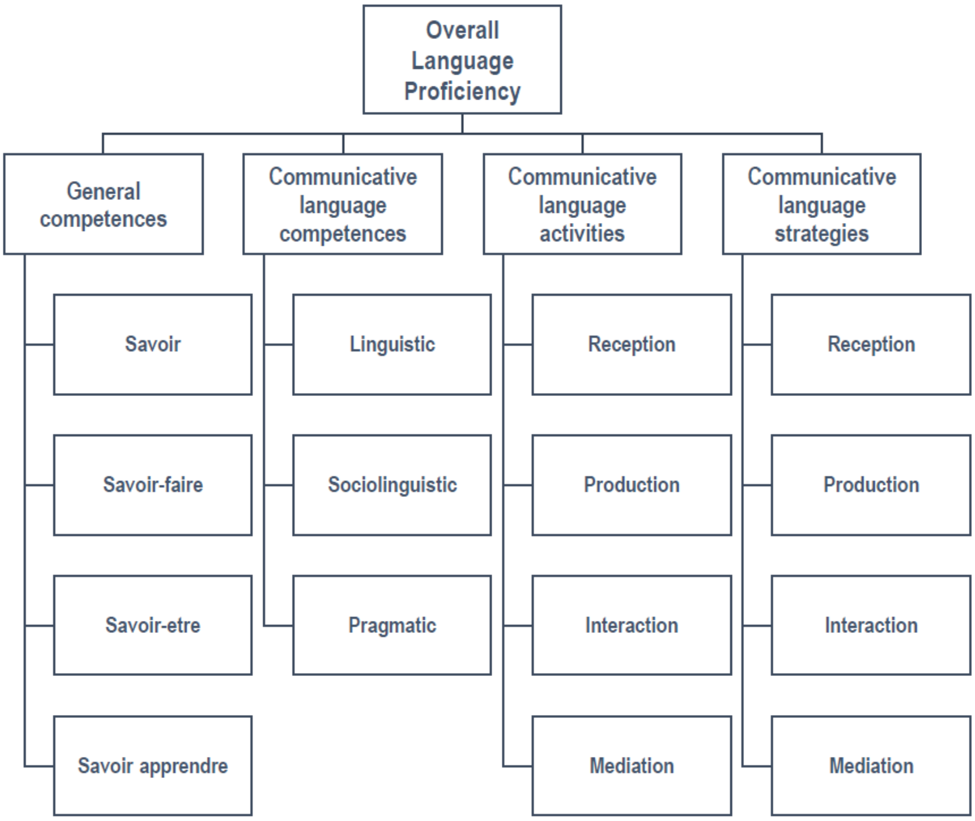


Figure 3 *The structure of the CEFR descriptive scheme*[7]

According to Council of Europe (2018: 28), such a view of the language learner and the language use and learning accords with the approach to teaching and learning suggested by the CEFR, which is that language learning should be directed towards enabling learners to act in real-life situations, expressing themselves and accomplishing tasks of different natures. It implies that the teaching and learning process is driven by action, that it is action-oriented. It also suggests planning backwards from learners' real life communicative needs, with consequent alignment between curriculum, teaching and assessment. Both the CEFR descriptive scheme and the action-oriented approach put the co-construction of meaning (through interaction) at the centre of the learning and teaching process.

The CEFR scheme is compatible with several approaches to second language learning, including the task-based approach (also known as communicative language teaching approach, CLT) (Council of Europe, 2018: 30). The CLT approach emphasises meaning-focused interaction in the target language, the choice of topics and activities that resemble real-life communication, the use of authentic texts and tasks, and a focus on the learning process itself (e.g. Wingate, 2018; cf. Nunan 1991; Mitchell 1994; Sauvignon, 2000). Aspects of the communicative approach appear to be suggested in the GCSE MFL curriculum (DFE, 2015) and used in GCSE MFL teaching (e.g. Bauckham, 2018; Wingate, 2018) although it is not entirely clear whether this is the dominant approach in all MFL classrooms in England.

Wingate (2018: 443) gives a useful history and summary of the curriculum and teaching approach in England in KS3 and KS4. Since MFL was included in 1992 as

[7] Taken from Council of Europe (2018: 30).

a foundation subject in the new National Curriculum (NC) framework for KS3 and KS4, the first policy document of the National Curriculum for Modern Foreign Languages (DES/WO 1991), as well as its subsequent versions, followed the CLT approach. According to this author, although CLT was not explicitly mentioned in the NC documents, this orientation was obvious in the educational purposes stated in the original policy document, and in the associated Programme of Study (PoS). The first of 8 educational purposes is 'to develop the ability to use the language effectively for purposes of practical communication' (DES/WO 1990: 3). Wingate cites Mitchell (2003: 18), who explains in reference to the 1999 version of the NC, that the PoS 'clearly encourage maximising learners' involvement in meaningful target language use'.

The new Department for Education subject content for reformed GCSE MFLs (DfE, ibid.: 3) lists the following as subject aims and learning outcomes, which should enable students to:

- develop their ability to communicate confidently and coherently with native speakers in speech and writing, conveying what they want to say with increasing accuracy
- express and develop thoughts and ideas spontaneously and fluently
- listen to and understand clearly articulated, standard speech at near normal speed
- deepen their knowledge about how language works and enrich their vocabulary in order for them to increase their independent use and understanding of extended language in a wide range of contexts
- acquire new knowledge, skills and ways of thinking through the ability to understand and respond to a rich range of authentic spoken and written material, adapted and abridged, as appropriate, including literary texts
- develop awareness and understanding of the culture and identity of the countries and communities where the language is spoken
- be encouraged to make appropriate links to other areas of the curriculum to enable bilingual and deeper learning, where the language may become a medium for constructing and applying knowledge
- develop language learning skills both for immediate use and to prepare them for further language study and use in school, higher education or in employment
- develop language strategies, including repair strategies

As with the previous versions of the NC, achievement of these goals would suggest development of communicative language competence, as well as use of the CLT approach. Therefore, the pedagogy of GCSE MFLs, and the approach in their associated assessments, should be compatible with the CEFR view of the language learner and language learning process, and thus not preclude a description of GCSE MFL performance standards and assessment standards in terms of the CEFR.

It should be noted, however, that available research suggests that current teaching methodologies at KS3 and KS4 may not be implementing the CLT approach in the way it was intended (e.g. Wingate, 2018, Bauckham, 2016). According to Bock (2002: 20, cited in Wingate, ibid.) the adaptation of CLT approach in the National curriculum for MFL has been accused of representing a narrow understanding of communicative competence and drawing 'on a rather selective interpretation' of the original principles (Block 2002: 20). This 'partial' and 'rather simplified version' (ibid.) has been blamed for over-emphasising speaking drills while at the same time failing to develop linguistic competence (Klapper 1997, 1998; Meiring and Norman 2001), knowledge about language, learner autonomy and intercultural competence (Pachler 2000). Mitchell and Martin (1997: 23) found in a study of French lessons in English secondary schools that 'learners were explicitly taught a curriculum consisting very largely of unanalysed phrases' which were 'memorised and rehearsed unaltered'.

According to Wingate (ibid.), some of this may be related to misconceptions about CLT in its strong version (cf. Swan, 1985) that instructed foreign language learning works in the same way as first language acquisition, and that learners would acquire grammatical structures implicitly from target language input. At the time of the NC's implementation, second language acquisition theory had recognised the need for 'focus on form' (Long 1991) alongside the focus on meaning (Wingate, ibid.: 444).

Based on a small-scale study in KS3 context, Wingate (ibid.) suggests that the teaching practices may now have shifted from the earlier CLT-orientation and may currently be dictated by the attainment targets that demand grammatical accuracy. While it is unclear whether a similar situation pertains to KS4 classrooms currently (although there are suggestions that this may be so, see Bauckham, 2016b), we believe it is important to bear in mind these indicators that MFL pedagogy in England may not be following the practices most widely recommended internationally.[8]

We do not believe that possible discrepancies between the notions of communicative language competence and language use as described in the CEFR, and the way communicative language competence and use may be understood, taught, and assessed at GCSE level in itself would invalidate an attempt to describe GCSE MFLs in terms of CEFR descriptors. We would argue that as long as the broad intention of the GCSE MFL curriculum and pedagogy is reasonably aligned to the CEFR – and this would appear to be the case as, for instance, they should "develop [learners'] ability to communicate confidently and coherently with native speakers in speech and writing, conveying what they want to say with increasing accuracy" – a description in terms of the CEFR may not only be appropriate, but also helpful.

However, we do believe that it is important to be aware of this context, as it may account for occasional disjoint between CEFR descriptors and GCSE assessments/performances that are observed in the linking. In addition, an awareness of these discrepancies could be helpful for improving both current

[8] Wingate (ibid.: 444) notes that although CLT has generally been regarded as an approach that motivates learners because it offers topic relevance and learner choice, current research suggests that this may not be the case with the in MFL classrooms in England. Various motivation studies carried out in the first ten years since the inception of the NCMFL (e.g. Chambers 1999; Graham 2002) revealed that MFL was the least popular subject and pupils found language lessons boring and repetitive. As Mitchell (2000: 288) explained, 'the curriculum may be too narrowly focused on pragmatic communicative goals, so that insufficient educational challenge is offered, with negative impact on pupil motivation'. Bartram (2005) found that pupils' attitudes towards learning French were negative because their use of language was limited to specific phrases prescribed for narrow communicative situations. In a review of the situation of language learning in English schools, commissioned by the government, Dearing and King (2007) criticised the lack of engaging curricular content and the fact that 'the present GCSE does not facilitate discussion, debates and writing about subjects that are of concern and interest to teenagers'. Macaro (2008) argued that many pupils lose motivation early on in KS3, because they are aware of a lack of progress and their inability to interact in the target language.

language pedagogy and assessment methods where appropriate, helping learners to achieve the goal of communicative language competence at the level appropriate for the phase of education at which they are.

# Method

## Overview

The approach in this study was guided by the recommended methods and procedures in the manual for relating language examination to the CEFR (Council of Europe, 2009) (henceforth, the Manual), and the updated descriptors from the companion volume (Council of Europe, 2018). The study was designed to provide empirical evidence for a link between performance and assessment standards of French, German and Spanish GCSE assessments at grades 9, 7 and 4, and the CEFR.[9]

Following the Manual, the study involved 5 stages:

1. familiarisation/training of participants,
2. content mapping (i.e., specification or relating the construct/content of the GCSE to the CEFR),
3. linking of performance standards for productive skills,
4. linking of assessment standards for receptive skills (including additional training/standardisation), and
5. empirical validation and evaluation.

Taken together, the results of stages 2-5 above should provide an indication of how GCSE performance and grading standards relate to the CEFR and its set of “can do” descriptors.

The figure below shows the sequence of key activities in the linking exercise. Each of the activities is described separately in more detail in the following sections.

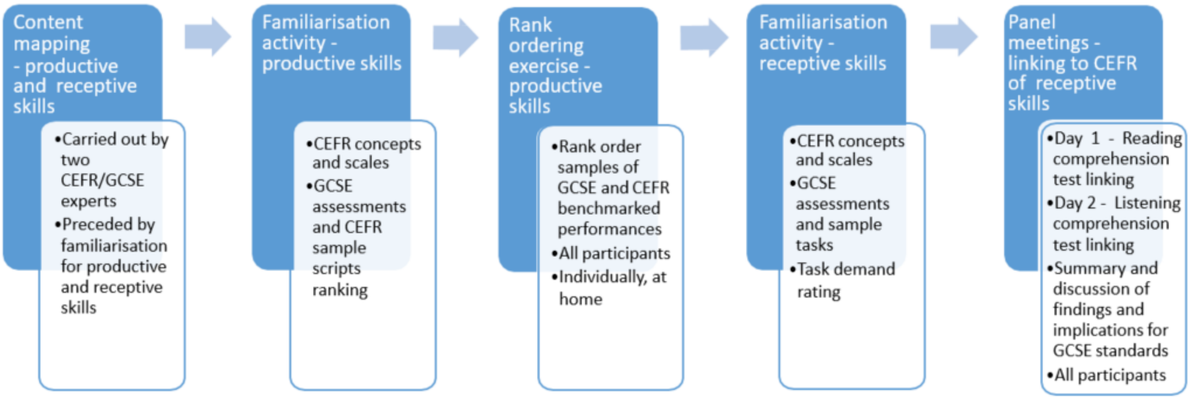

Content mapping - productive and receptive skills
•Carried out by two CEFR/GCSE experts
•Preceded by familiarisation for productive and receptive skills
Familiarisation activity - productive skills
•CEFR concepts and scales
•GCSE assessments and CEFR sample scripts ranking
Rank ordering exercise - productive skills
•Rank order samples of GCSE and CEFR benchmarked performances
•All participants
•Individually, at home
Familiarisation activity - receptive skills
•CEFR concepts and scales
•GCSE assessments and sample tasks
•Task demand rating
Panel meetings - linking to CEFR of receptive skills
•Day 1 - Reading comprehension test linking
•Day 2 - Listening comprehension test linking
•Summary and discussion of findings and implications for GCSE standards
•All participants

Figure 4 *Sequence of activities in the linking exercise*

[9] Note that work from higher tier only was considered for grade 4.

# Specifications

Three specifications with larger entries from 2 exam boards were chosen for the study:

- AQA GCSE French (8658)
- AQA GCSE German (8668)
- Pearson GCSE Spanish (1SP0)

All of these were new, reformed GCSE specifications developed for first assessment in summer 2018. Therefore, only assessment materials from the June 2018 examination session were available for the study. Only work from the higher tier was considered for grade 4. The table below shows maximum marks for each specification and paper.

Table 4 *Maximum mark for specifications and papers*

| Specification | Writing | Speaking | Reading | Listening | Total |
|---|---|---|---|---|---|
| French | 60 | 60 | 60 | 50 | 230 |
| German | 60 | 60 | 60 | 50 | 230 |
| Spanish | 60 | 70 | 50 | 50 | 230 |

# Participants

For each language, panels of 13 experts were recruited to participate. We endeavoured to recruit participants who had at least some familiarity with the CEFR. However, this was not possible in all cases. Nevertheless, the majority of the participants did have some relevant CEFR experience.

Each of the 3 panels consisted of: HE linguists, staff from international testing organisations (Institut Francais, Alliance Français, Goethe Institut and Instituto Cervantes), and Ofqual subject experts, all with reasonable experience or specialism in the CEFR; A level MFL teachers from both state and independent schools, most with some familiarity with the CEFR; representatives of subject associations; and representatives of exam boards. The participants from the last 2 groups did not necessarily have direct experience of using the CEFR.

Table 5 *Breakdown of panellist background/role by panel*

| Role | French | German | Spanish |
|---|---|---|---|
| HE experts | 3 | 5 | 5 |
| International testing organisations experts | 3 | 1 | 1 |
| A level teachers | 4 | 4 | 4 |
| Ofqual subject experts | 1 | 1 | 1 |
| Subject association reps | 1 | 1 | 1 |
| Examination board reps | 1 | 1 | 1 |

HE participants were recruited via contacts collated for a previous study (Curcin and Black, 2018). On this occasion, however, participation was conditional on practical experience with the CEFR.

The participants from international testing organisations were contacted via institutional email addresses in the first instance. The relevant institutions then chose the most suitable person with relevant CEFR experience, who took part in the study.

A level teachers were recruited by initially contacting administration offices of all state and independent secondary schools with more than 10 A level candidates in 2018. Again, participation ideally required some degree of familiarity with the CEFR.

Table 6 *Breakdown of A level teacher school type and CEFR familiarity by panel*

| Panel | Participant ID | School type | CEFR familiarity |
|---|---|---|---|
| French | J05 | Sixth Form College | N |
| | J07 | Grammar school | Y |
| | J31 | Academy | Y |
| | J33 | Independent | Y |
| German | J04 | Independent | Y |
| | J08 | Academy | Y |
| | J12 | Grammar school | Y |
| | J13 | Sixth Form College | N |
| Spanish | J03 | Grammar school | Y |
| | J10 | Grammar school | Y |
| | J15 | Sixth Form College | Y |
| | J16 | Independent | Y |

Ofqual subject experts were recruited by sending invitations to participate to all experts on the Ofqual list with relevant subject expertise. One of the requirements for participation was a reasonable practical experience of using the CEFR.

Subject associations and exam boards were invited to send a representative for each language, where possible with some familiarity with the CEFR. The representatives of exam boards were either examiners or subject experts. They were allocated to panels such that the representatives came from a different exam board from that whose specification was the focus of the panel. Thus, a WJEC representative attended the French panel, a Pearson representative attended the German panel, and an AQA representative attended the Spanish panel.

At the start of their online familiarisation, the participants were asked several questions about their experience of and attitudes towards the CEFR. Figure 5 shows a breakdown of participant CEFR familiarity levels prior to familiarisation by panel. The charts show a similar pattern across the 3 languages, with the majority of the participants having an interest in the CEFR, some theoretical or academic knowledge of it and some practical experience of using it in the context of teaching and marking. Over half of the participants in each panel had some experience of using the CEFR in the test or resource development. Except for the Spanish panel, few participants had experience of using the CEFR in the context of teaching English as a foreign language, while the majority in every panel had experience of using it in the context of teaching the target language of the panel as a foreign language. One or 2 participants had experience of the CEFR solely based on teaching English as a foreign language.

Figure 6 shows participants’ attitudes towards the CEFR for each panel. With very few exceptions, the participants either strongly agreed or agreed with the way the CEFR describes differences in learner ability levels. Similar attitude was expressed towards the statement that understanding GCSE standards in relation to the CEFR may be helpful.

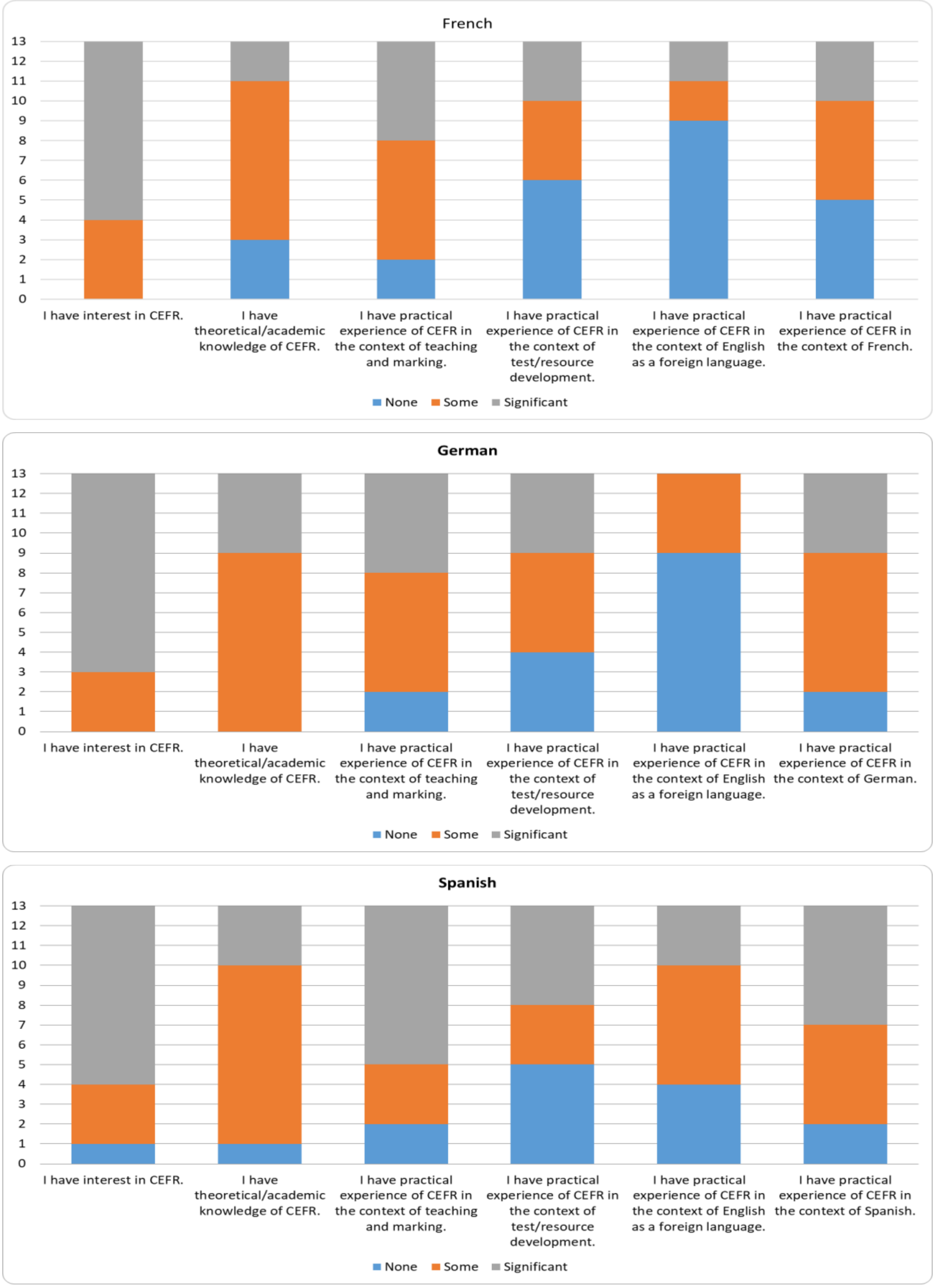


Figure 5 *Nature of participants' experience with the CEFR*

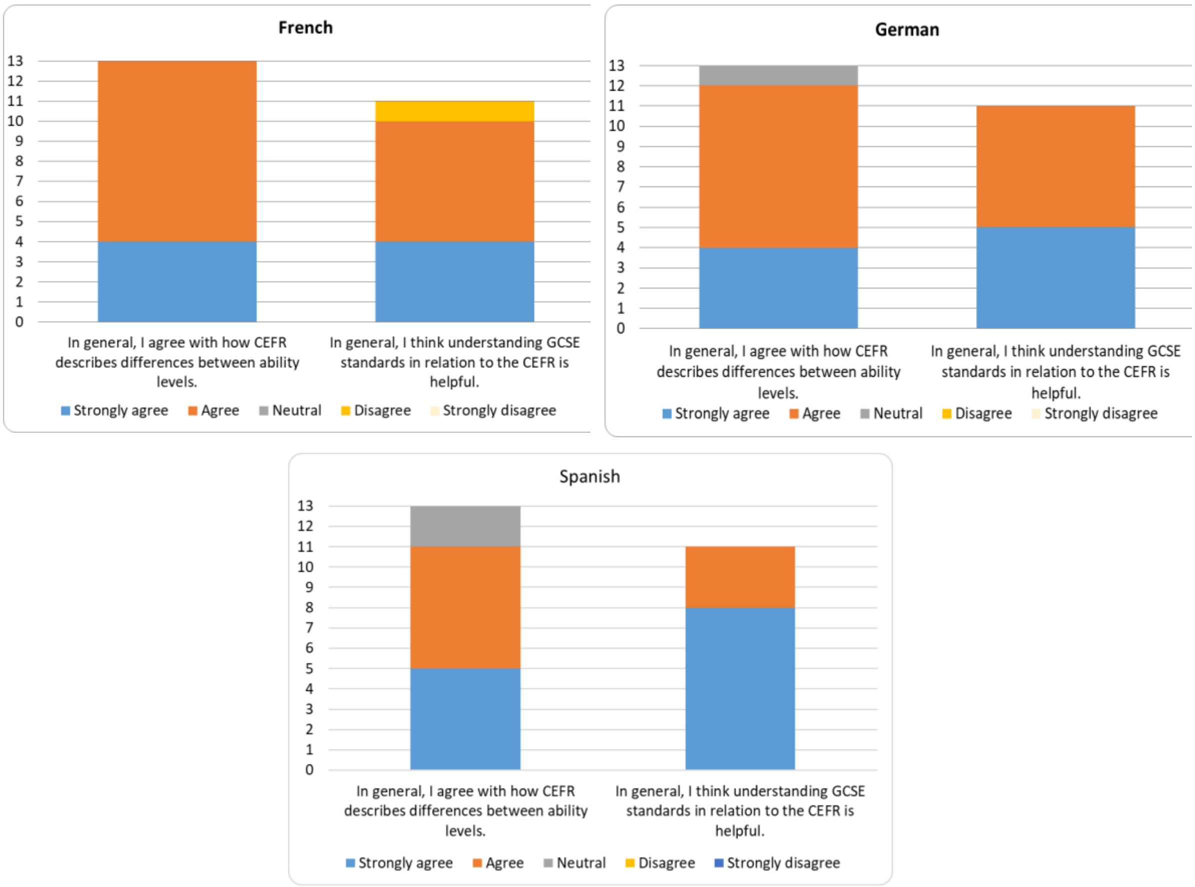


Figure 6 *Participants' attitudes to the CEFR and its use in understanding GCSE standards*

As part of their online familiarisation for the receptive skills, the participants were asked about their experience of writing reading or listening comprehension tasks in either panel target language or another language, as well as about their experience of writing reading or listening comprehension tasks targeted at specific CEFR levels. Figure 7 shows that the majority of participants in each panel stated that they had at least some experience in each of these domains. Only one participant in the German panel and one in Spanish had some experience of standard setting for language tests.

Given that the starting point for the majority of the participants was some familiarity and practical experience of using the CEFR, it was hoped that further familiarisation and training would help to get everyone to a level where they can usefully contribute to the linking study. In particular, further opportunity for discussion of the relevant CEFR scales in relation to the standard linking method used for listening and reading comprehension assessments, was provided at the start of each standard linking meeting.

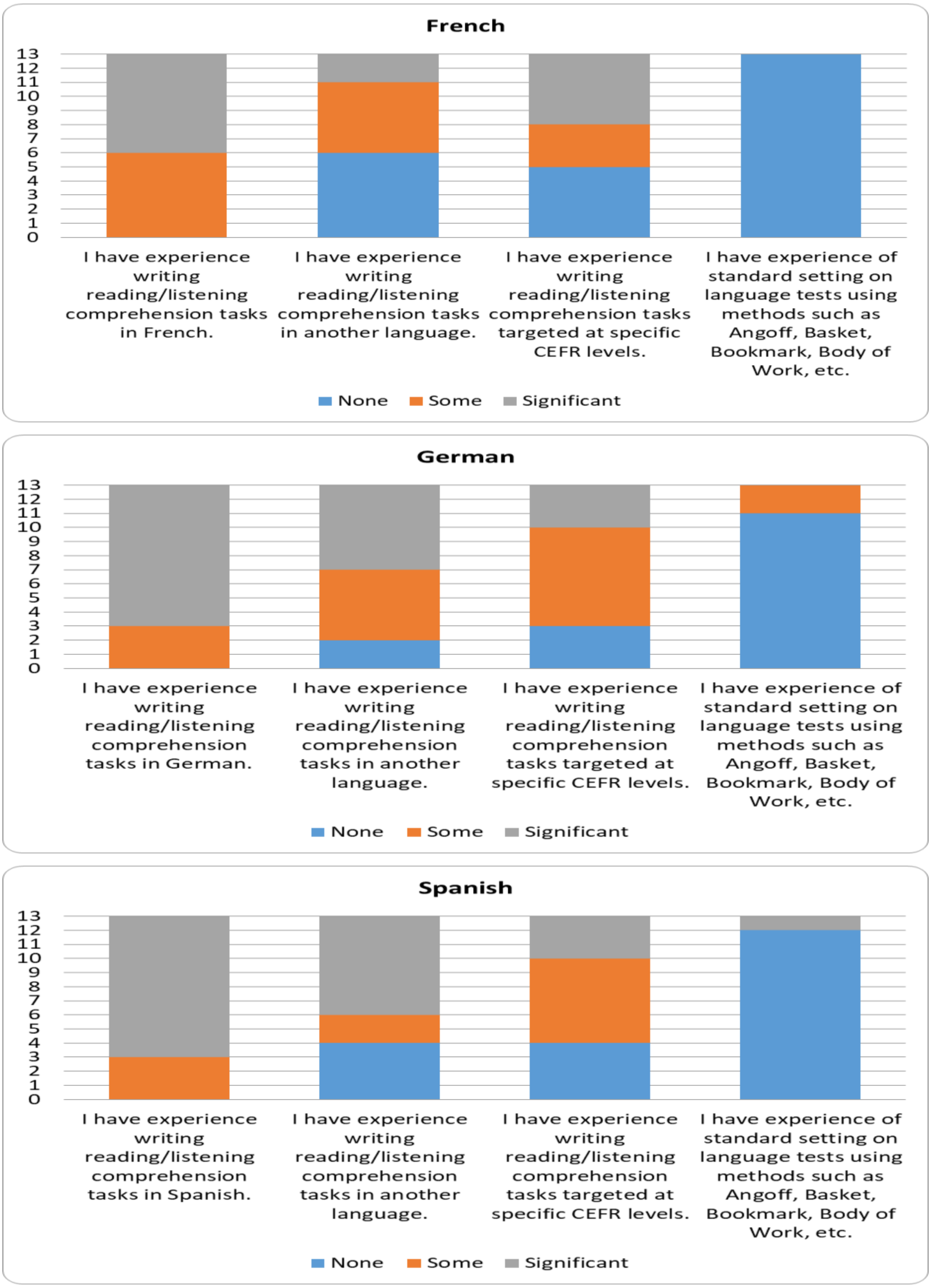


Figure 7 *Experience of writing reading/listening comprehension test items and standard setting*

## Familiarisation and training

Prior to undertaking any of the main activities in the study, the participants were provided with familiarisation and training to ensure reasonable individual and common understanding of the relevant aspects of the CEFR and of the GCSE assessments. This aimed to ensure the integrity and quality of panellists' judgements.

Separate familiarisation and training activities were created for productive skills and receptive skills. Participants were contracted to complete familiarisation activities in half a day for productive skills and half a day for receptive skills.

The majority of familiarisation and training activities were conducted online, using a survey tool set up with a range of activities. Some of the activities required reading of materials provided outside of the training tool. These had been provided in hard copy. Some activities involved ranking or rating of performances and/or test questions, which were accessed electronically via the links provided within the training tool. The contents of each training tool, alongside the various documents and CEFR scales provided to the participants, are presented in Appendix B and C.

The participants who took part in content mapping were provided with the familiarisation and training activities before they carried out the content mapping activities. They completed familiarisation for productive skills first, followed by content mapping for productive skills. After this, they completed familiarisation for receptive skills, followed by content mapping for receptive skills.

The rest of the participants were first provided with familiarisation for productive skills, following which they carried out the rank ordering for productive skills (see below). Given the constraints of participant availability, it was not possible to arrange for separate face-to-face training and discussion sessions ahead of the rank ordering exercise. However, it was hoped that the intuitive nature of the rank ordering task, which is typically conducted individually from home, and helps to cancel out systematic biases and severity/leniency effects in judgements, would have made up for absence of face-to-face training (cf. Black and Bramley, 2008; Curcin and Black, in prep; Jones, 2009).

Familiarisation for productive skills included the following key aspects:

- reading of excerpts from the CEFR document (ibid.) which briefly described what the CEFR is, its conceptualisation of language ability, what illustrative descriptors are and how to read them
- familiarisation with the global CEFR scale,sorting individual CEFR descriptors from the CEFR global scale into levels
- self-assessment of participants' own CEFR level using CEFR descriptors
- familiarisation with overall written and spoken production and interaction and mediation CEFR scales
- consideration of examples of written and spoken performances with known CEFR levels and deciding on key features that distinguish between performances at different CEFR levels
- familiarisation with GCSE specifications and assessment materials, including sketching answers to each question paper; familiarisation with processes of marking and grading in GCSEs

- familiarisation with rank ordering written and spoken performances, and
- exercises in ranking written and spoken performances

Initial familiarisation for receptive skills was provided about a week ahead of the panel meetings during which the linking of assessment standards for receptive skills was conducted. It was conducted individually, from home, using the training tool provided. Further opportunity for discussion of the relevant CEFR scales in relation to the method used for linking the standards, as well as in relation to use of the CEFR in standard linking exercises, was provided at the start of each standard linking session.

Initial familiarisation for receptive skills included the following key aspects:

- familiarisation with the concepts of task[10] and question demand vs. difficulty and the difference between these in the context of standard setting, including reading excerpts from the CEFR document (2001) about features that can affect comprehension task difficulty
- consideration of which aspects of text/audio and question demand in a test could be manipulated to change the level of demand, with particular reference to demand features of translation tasks
- familiarisation with threshold and "plus" level illustrative descriptors for comprehension and what it means for a learner to be at a threshold level familiarisation with overall reading and listening comprehension scales
- familiarisation with additional reading and listening comprehension scales
    - reading correspondence
    - reading for orientation
    - reading for information and argument
    - reading instructions
    - identifying cues and inferring (spoken and written)
    - understanding conversation between other speakers
    - listening to announcements and instructions
    - listening to audio media and recordings
- familiarisation with aspects of general linguistic competence and some of the relevant scales (general linguistic range, vocabulary range, grammatical accuracy, flexibility)
- consideration of the most salient aspects that distinguish between tasks targeted at different CEFR levels, using exemplar tasks with known CEFR levels
- familiarisation with rating reading and listening comprehension tasks in relation to the key question that was going to be asked during standard linking meetings ("Which is the first CEFR level describing learners who can answer this task correctly?")
- familiarisation with GCSE specifications and assessment materials for reception, including sketching answers to each question paper, and

[10] See more details on this in the description of the standard linking procedure below.

- exercises in rating CEFR exemplars and GCSE reading and listening comprehension tasks in relation to the abovementioned question

In addition to these activities, at the start of each panel meeting further familiarisation was conducted including,

- clarification of any concepts and issues from initial familiarisation
- further, more detailed description of the standard linking process
- discussion of the ratings collated from the survey tool (see graphs at the end of this section), including a wider discussion about features that contribute to text and question demand, further discussion of how to answer the key standard linking question, and what constitutes threshold performance at different CEFR levels

The CEFR benchmark performances, as well as benchmarked reading and listening tasks, were obtained from the Council of Europe website from the following links:

- Writing: https://www.coe.int/en/web/portfolio/reading-listening-and-writing[11]
- Speaking: http://www.ciep.fr/ressources/ouvrages-cederoms-consacres-a-levaluation-certifications/dvd-productions-orales-illustrant-les-6-niveaux-cecrl
- Reading: https://www.coe.int/en/web/common-european-framework-reference-languages/reading-comprehension
- Listening: https://www.coe.int/en/web/common-european-framework-reference-languages/listening-comprehension

The CEFR benchmark speaking performances were presented as audio rather than video files to make them more comparable with GCSE speaking performances, which were only available as audio files.

While some aspects of the training were “open book”, the participants were encouraged to approach the tasks professionally and were given sufficient time to enable them to devote due attention to all activities. The figures below show a breakdown of participants’ evaluation of the effectiveness of different aspects of familiarisation after completing the activities. The results suggest that the participants generally considered the activities effective in helping them become confident in using the CEFR in relation to both the productive skills performances and receptive skills assessment demands. The participants were slightly more likely to disagree with the statements about the effectiveness of training for receptive skills (Figure 9).

[11] Except for Spanish writing exemplars, which were obtained from Instituto Cervantes directly (also available at https://londres.cervantes.es/en/default.shtm).

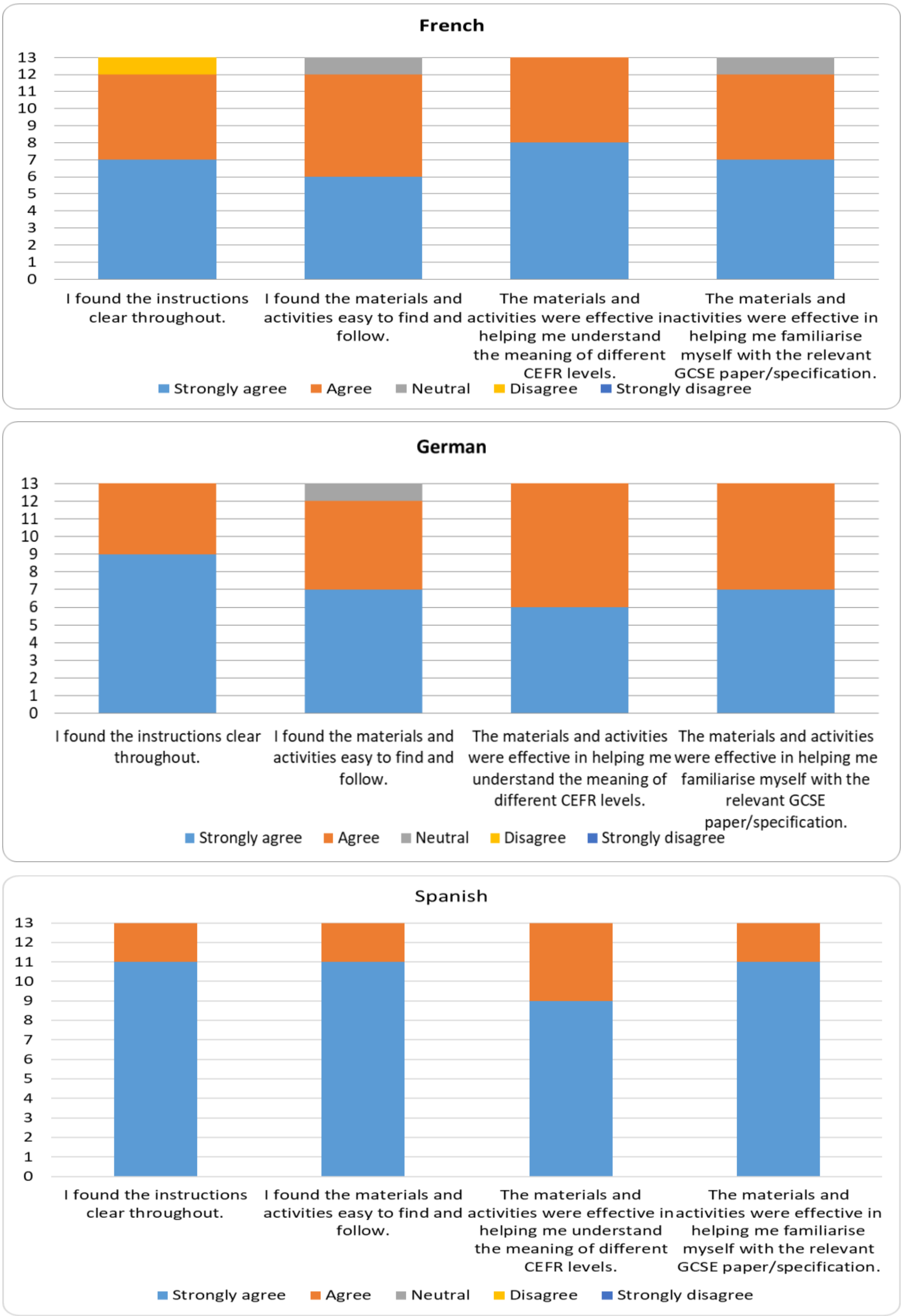


Figure 8 *Training evaluation – productive skills*

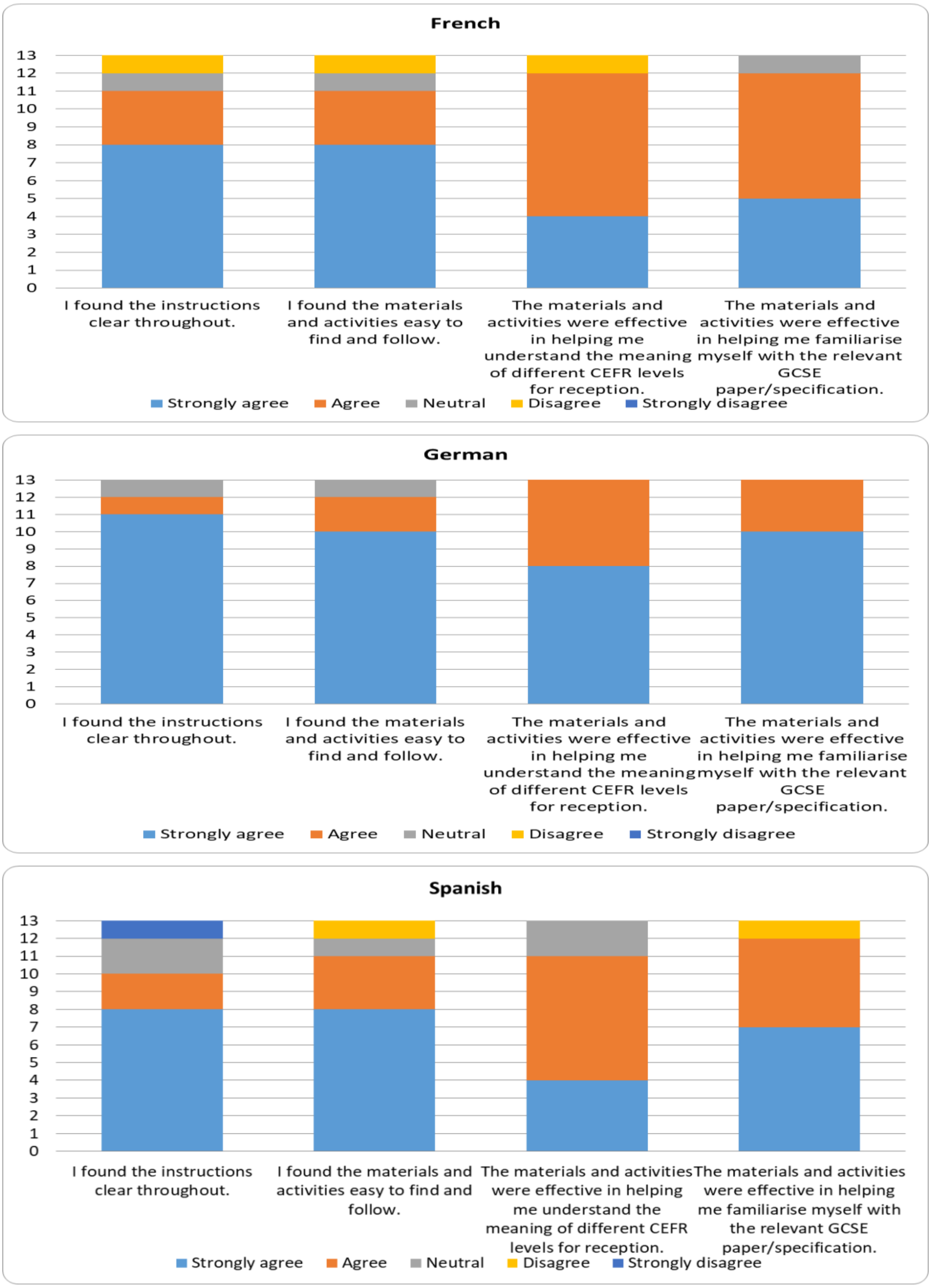


Figure 9 *Training evaluation – receptive skills*

The participants were also asked about how confident they were in understanding the distinction between CEFR levels at the end of the training. Figure 10 shows that, while the majority were either very confident or confident, they appeared less confident about the training affording them an understanding of the difference between CEFR levels for receptive skills. This was unsurprising as it is well known that determining demand levels of questions or tasks, especially for specific ability levels of learners, is very difficult. For this reason, we allowed additional time during panel meetings for group discussion and to clarify any issues and misconceptions.

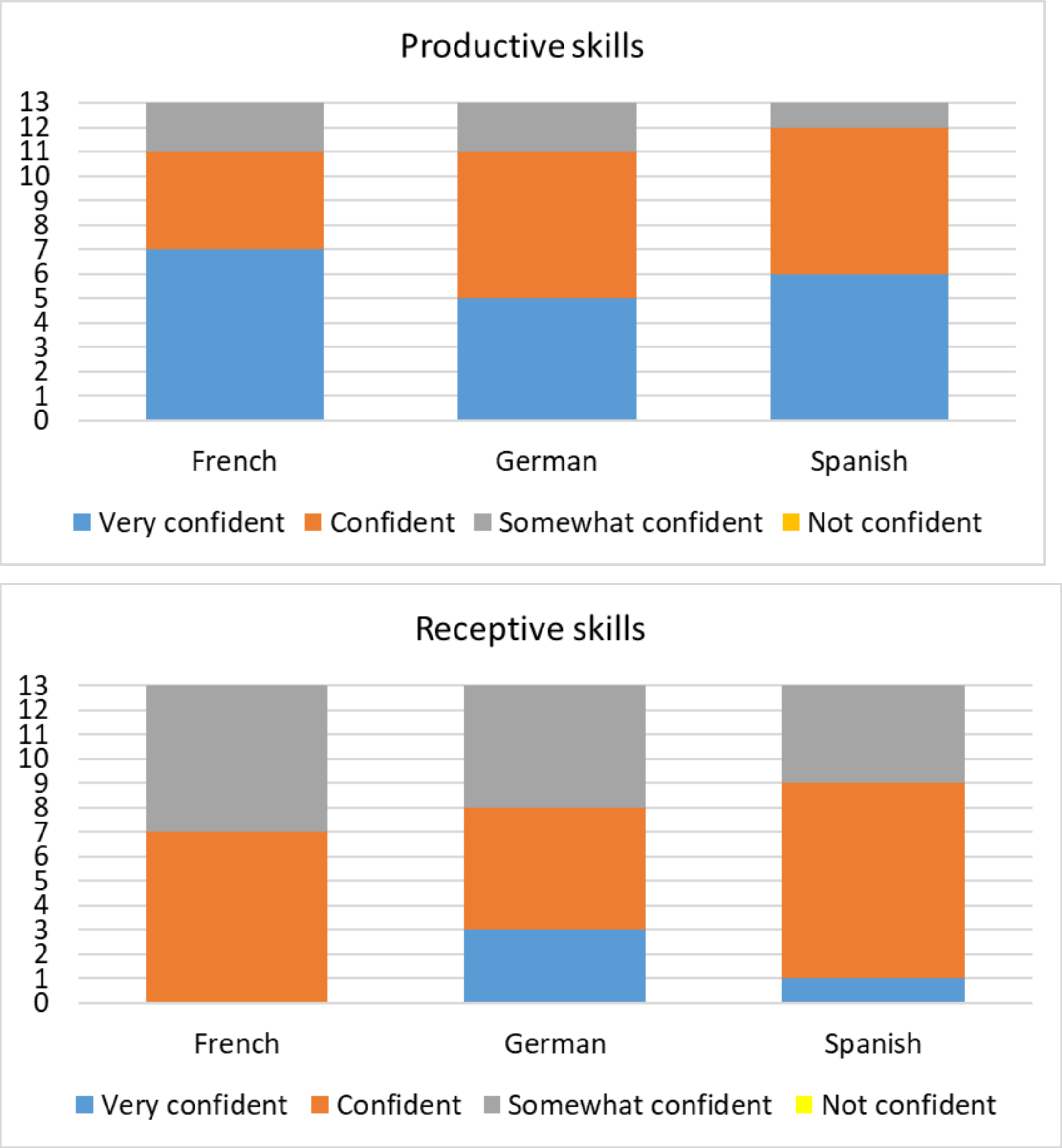


Figure 10 *Confidence in understanding the distinction between CEFR levels at the end of the training*

The figures below present distributions of ratings that the participants gave during initial familiarisation to the CEFR exemplar tasks, for which they did not know the actual CEFR level rating (denoted by the large label on each chart). These exemplars and ratings distributions were further discussed at the start of each panel meeting to help bring the participants closer together in terms of their common understanding of the CEFR scales as applied to reading and listening comprehension tasks.

Given that these ratings are based only on what the participants already knew or learnt from our online familiarisation, and prior to the additional time spent familiarising and standardising the participants at the start of each panel session, these distributions can be characterised as reasonably good in terms of agreement levels and accuracy of rating compared to the actual CEFR level of each example. This provides further evidence of a reasonable level of understanding on the part of our participants of CEFR rating scales and categories, even before additional training and further discussion in the meetings.

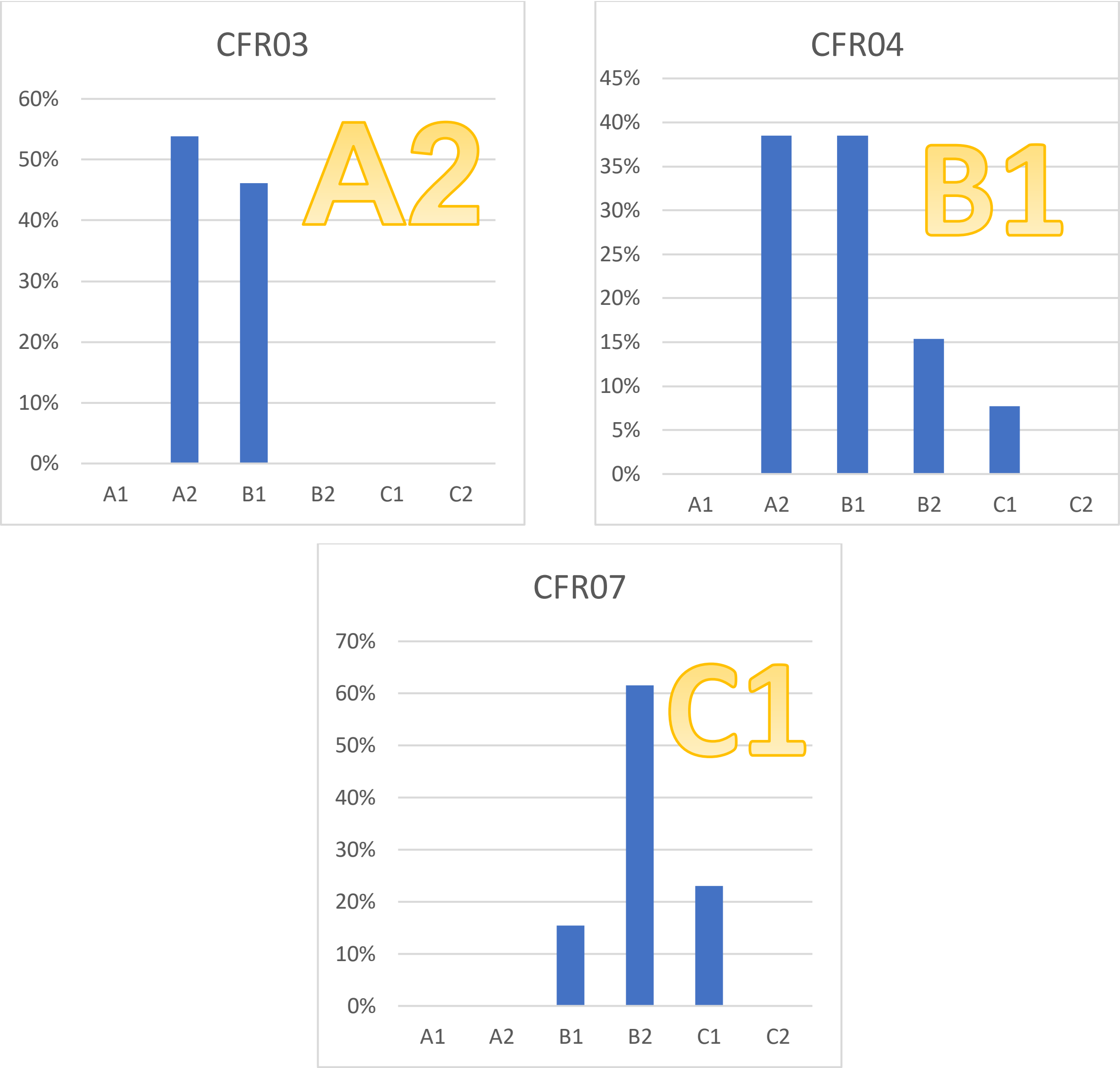


Figure 11 *Familiarisation ratings distribution of CEFR exemplars – French reading*

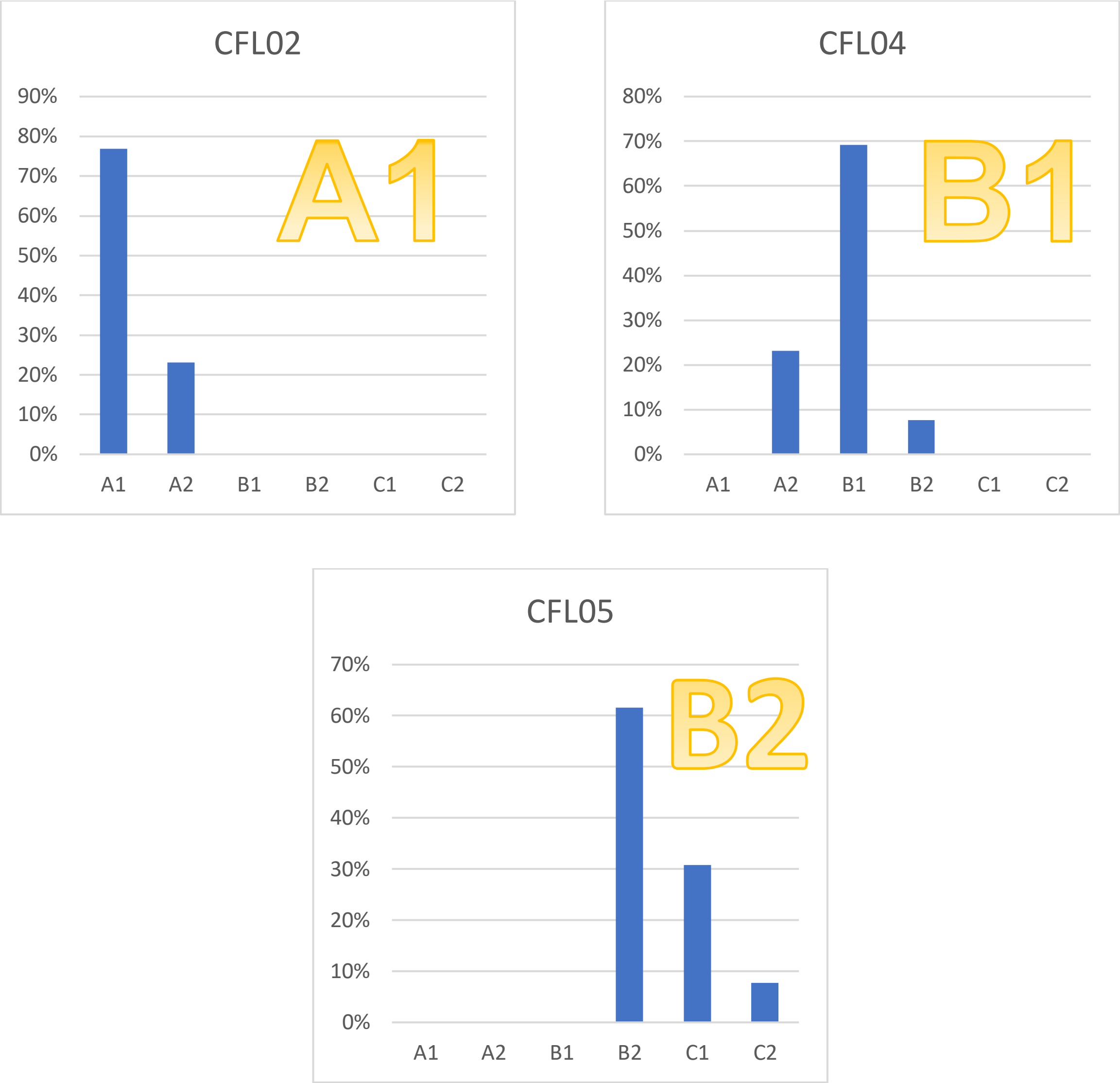
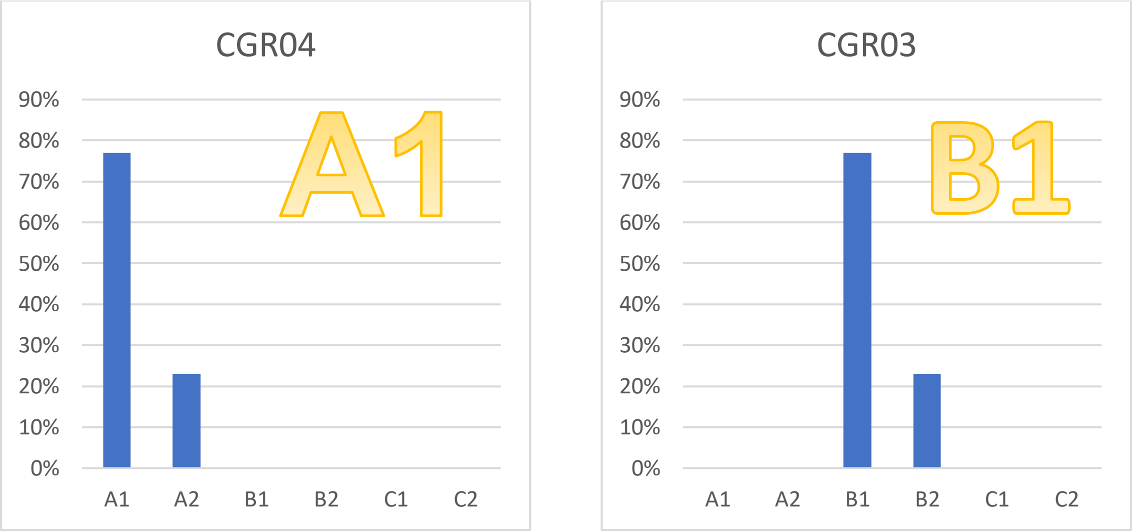


Figure 12 *Familiarisation ratings distribution of CEFR exemplars – French listening*



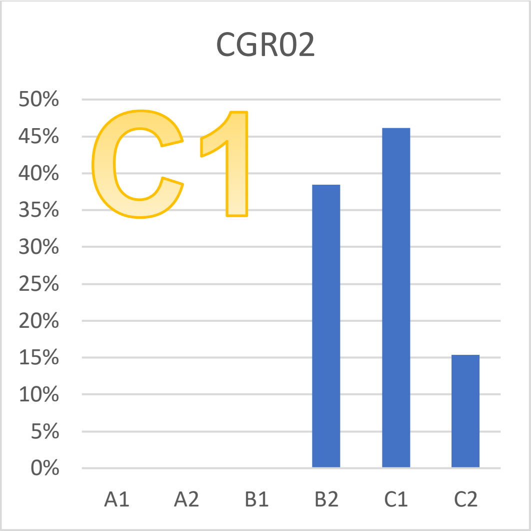


Figure 13 *Familiarisation ratings distribution of CEFR exemplars – German reading*

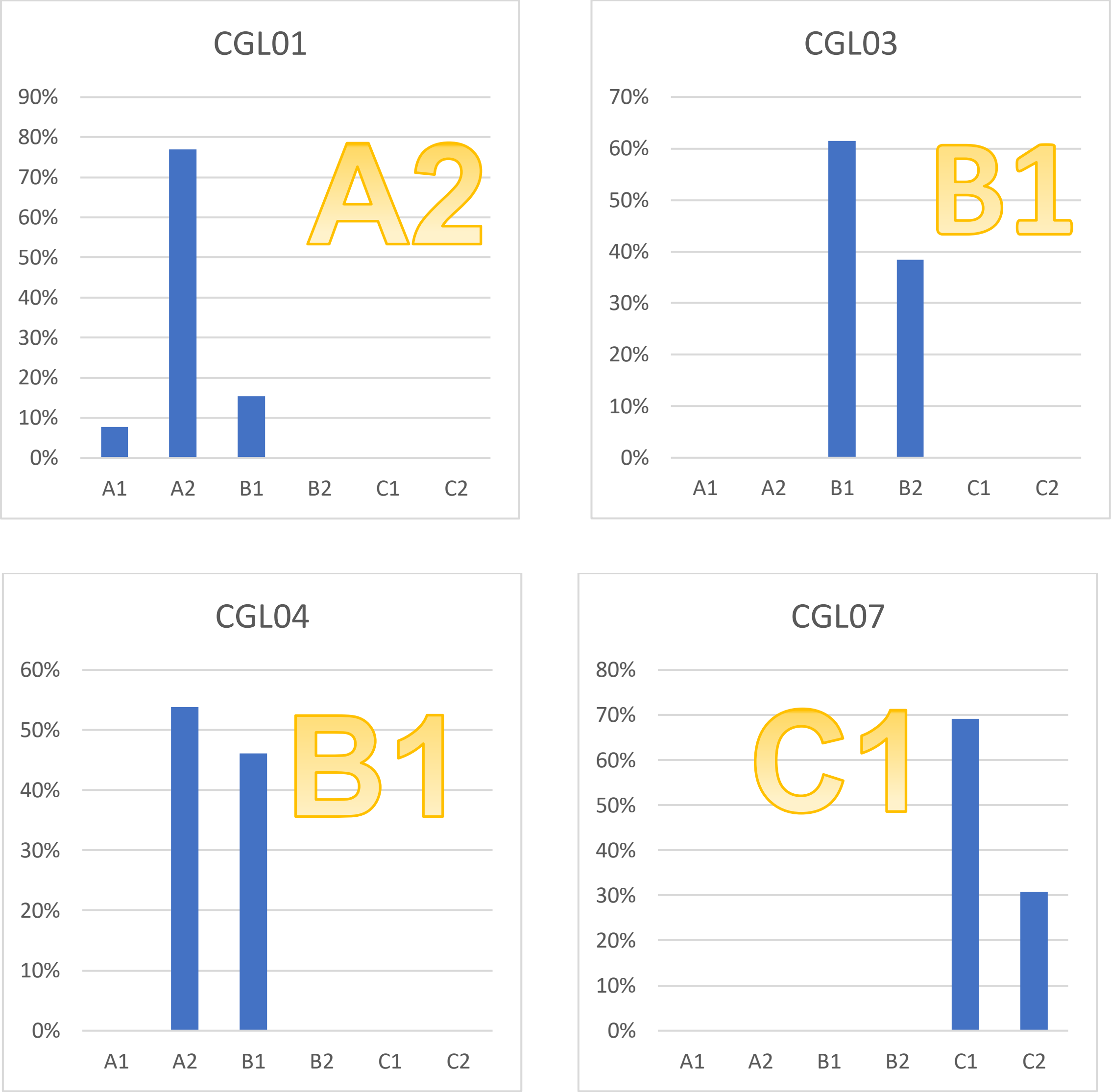


Figure 14 *Familiarisation ratings distribution of CEFR exemplars – German listening*

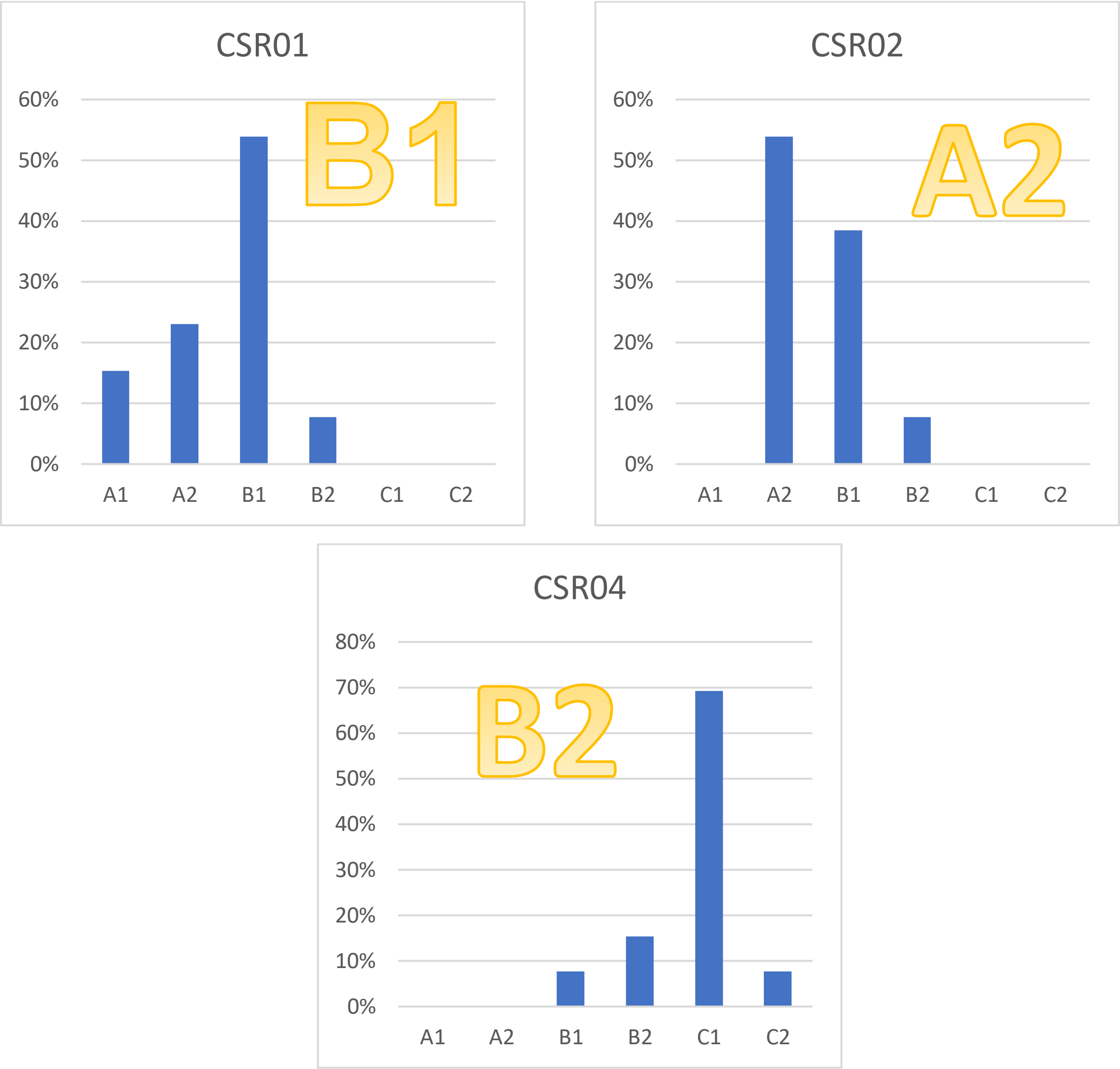


Figure 15 *Familiarisation ratings distribution of CEFR exemplars – Spanish reading*

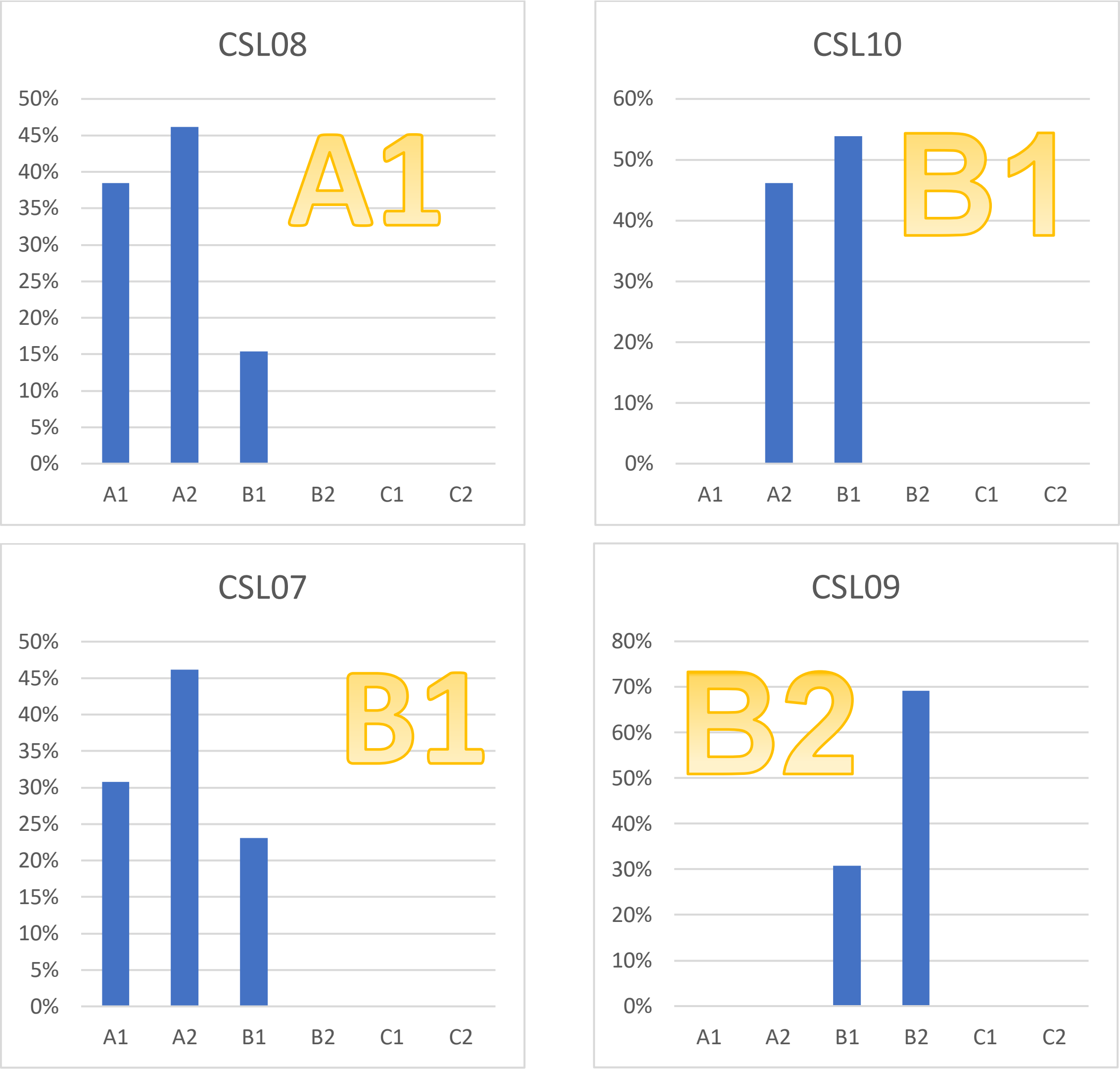


Figure 16 *Familiarisation ratings distribution of CEFR exemplars – Spanish listening*

## Content mapping

This exercise involved relating the content of the GCSE specifications and summer 2018 assessments, and their intended demand, to CEFR levels. This task was carried out by one CEFR expert and one GCSE subject expert per language independently of one another initially. Content mapping took about 2 working days per participant. This was followed by a discussion meeting, one for each language, facilitated by an Ofqual researcher, in order to come to an agreement on any discrepancies in individual views. The final CEFR ratings were confirmed at the end of this meeting.

This exercise provides the initial, tentative link to the CEFR, but is usually not considered detailed enough nor sufficient to provide enough evidence to support the linking argument (see e.g. the Manual, ibid.). The end product of this exercise is a claim of a degree of linking of GCSE MFLs to the CEFR based on profiling the examination in relation to CEFR categories and levels. In addition, in the context of the current study, this mapping helped to understand the extent to which it is appropriate to link the GCSE grade scale to the CEFR on the basis of alignment of the 2 constructs (i.e. coverage, intended demand, etc.). This is particularly important given that GCSE MFL assessments were not developed explicitly with reference to the CEFR.

In the instructions provided to the participants, it was emphasised that content mapping and profiling of the specification and the assessment according to the CEFR levels should be done bearing in mind the intended purpose, coverage and demand of the exam, rather than with reference to what a current cohort of students might be able to do in the exam. This is because, for example, if an exam is designed to target CEFR levels A1 to B1, it would still be possible that the performances that it elicits from candidates are all at A2 level or below. This could mean that the exam is not targeted well enough for its intended population even though the level of the exam itself might be well aligned with its purpose.

Alongside the instructions, further materials were provided to the participants to help ensure a thorough understanding of the relevant aspects to consider in content mapping.[12] These included:

- the CEFR full text (ibid.)
- the companion volume (ibid.)
- reference scales for production, interaction, reception and mediation to use as relevant
- excerpts from the CEFR text regarding a classification of communication themes
- table 5 from the CEFR text detailing descriptive categories for external context of language use
- table A3 from the Manual, detailing relevant qualitative factors for reception,
- tables A4 and A5 from the Manual, detailing relevant qualitative factors for spoken interaction and production respectively

We only asked the participants to profile the content and intended demand of the specification and assessment, and did not ask them to consider aspects such as test development processes, assessment functioning, etc., which are sometimes included in the specification phase. These aspects are considered as part of standard examination board processes as well as in Ofqual monitoring and other research studies, and were thus not the focus of this study.

Materials used for content mapping were adapted from the Manual (ibid.). The materials included:

- Form A1 - general examination description
- Form A9 – listening comprehension
- Form A10 – reading comprehension
- Form A11 – spoken interaction
- Form A12 – written interaction
- Form A13 – spoken production
- Form A14 – written production

[12] Most other materials that the participants were provided are included in Appendix D.

- Form A15 – integrated skills combinations
- Forms A17 and A18 – spoken and written mediation
- Form A19 – communicative language competence in reception
- Forms A20-A22 - communicative competence in interaction and production
- Adapted form A24 – confirmed estimation by skill/component[13]
- GCSE specifications, assessment materials (including audio materials) and mark schemes

In the instructions document, the participants were asked to bear in mind that some of the forms may not be entirely relevant for GCSE MFL examinations, e.g. those relating to mediation or integrated skills.[14] In particular, we asked the participants to only consider integrated skills at high level, rather than filling in a form for each combination separately. We also asked them to only consider written mediation, and to request more detailed forms if appropriate. Also, we did not provide forms that relate to communicative language competence specific to mediation and the participants were asked to request these if deemed useful and appropriate. We asked the participants to advise about which detailed reference scales it would be useful to include during participant training and for other activities. Finally, the participants were asked to consider whether there are any significant obstacles in terms of content, purpose, or any other reasons why an attempt to link GCSE and CEFR standards would be inappropriate.

## Rank ordering of written and spoken performances

The second phase of the study was to relate the performance standards of productive skills (writing and speaking) at grades 9, 7 and 4 to the CEFR. We used the rank ordering method to do this, as previously advocated by Jones (e.g. 2009), separately for each skill. The rank ordering task was completed by all the participants. This was supplemented by obtaining ratings of the CEFR benchmarks used in the rank ordering exercise from the panel members at the end of the standard-linking panels. These ratings were collected in order to inform the analysis of the rank ordering data.

### *Rationale for the method*

Rank ordering is a technique for capturing expert judgement. It has been used considerably in the assessment context in the last decade. Previous research exercises have found that rank ordering is a valid method for comparing performance standards, for example, between examinations from different years (Black and Bramley, 2008; Curcin and Black, in prep.). Jones (ibid.) showed that the method can be used to replicate the results of panel ratings using CEFR descriptor scales. This technique would therefore allow us to map the CEFR levels and GCSE grades to the same common scale in order that we can understand the relationship between the two.

[13] The forms are included in Appendix E.

[14] Integrated skills involve a mixture of reception and production, for instance, listening to a text and answering questions, and then using the information gained to make a summary.

In this method, a sample of candidate performances (scripts) from 2 or more examinations are rank ordered by multiple judges (e.g., examiners, subject experts). These rankings are then combined and analysed using the Rasch model to place each script onto a single scale of quality. The theory underlying the rank ordering method is Thurstone's law of comparative judgement (Thurstone, 1927). Given that the judgements collected through the rank ordering method are subsequently analysed using the Rasch model, which allows for non-randomly missing data, it is possible to create judgment allocation designs that are sufficiently sparse to be feasibly implemented in practice, while being sufficiently large-scale to result in reasonably reliable estimates of the scale of interest (e.g. the script quality scale).

Given the naturally intuitive task of comparing performances to one another, as well as practical considerations, it was decided to conduct familiarisation and training for productive skills, as well as the main rank ordering tasks, online, as this allowed the participants to work remotely. The rank ordering method, by design, cancels out systematic individual participant biases and severity/leniency effects in judgements. In addition, it maximises the reliability of the perceived performance quality scale by enabling each performance to be seen by multiple judges in a fairly economical judging design. Furthermore, the method involves comparing performances with one another, and putting them in a rank order, rather than rating them with specific CEFR categories. Therefore, while it does require familiarity with the basic notions and approach inherent in CEFR descriptors, it does not require detailed knowledge of the CEFR scales.

## *Procedure*

In the current study, each participant rank ordered, in terms of overall quality, a series of GCSE scripts (at grades 9, 7 and 4) which were interspersed with performances from candidates in the same language which had previously been empirically benchmarked on the CEFR scale (from a range of different international exams, rather than from GCSEs). The participants did not know the CEFR levels of these benchmark examples nor did they know the marks or grades of the GCSE scripts. We asked the participants to take into account aspects of linguistic ability that are valued in the CEFR and its reference scales when comparing these performances.

The result of this exercise was a common script quality scale for the GCSE scripts and the CEFR benchmarks. By looking at how the grades of the GCSE scripts and the levels of the CEFR benchmark script are distributed on this quality scale, we can relate the performance standards at different GCSE grades to the CEFR scale.

The rank ordering exercises were conducted for writing and speaking separately. For each skill, 3 GCSE scripts on each of the boundary marks for grades 9, 7 and 4 and 2 mark points below each grade boundary were included in the rank ordering exercise. The CEFR benchmarked exemplars at levels A1, A2, B1, B2 and C1 were also included[15]. All marks and annotations, as well as indications of the CEFR level were removed from the scripts.

Samples of GCSE scripts were obtained from the relevant examination boards. The scripts in the sample were randomly chosen from among the best fitting scripts

[15] C1 for Spanish speaking was not included due to a labelling error, resulting in double the number of comparisons for the B1 exemplar. This is not a significant problem as C1 performances were not expected to be perceived to be close in quality to any GCSE grade boundary scripts.

based on a Rasch partial credit model item level analysis (cf. Raikes, Scorey and Shiell, 2008; Curcin and Black, in prep.).

The CEFR benchmark performances were obtained from the Council of Europe website from the links below (except for Spanish writing, which were obtained from Instituto Cervantes). For the rank ordering, we used those exemplars that had not already been used in familiarisation activities.

- Writing: https://www.coe.int/en/web/portfolio/reading-listening-and-writing
- Speaking: http://www.ciep.fr/ressources/ouvrages-cederoms-consacres-a-levaluation-certifications/dvd-productions-orales-illustrant-les-6-niveaux-cecrl

The judges were contracted for half a day per component for this exercise. Each judge saw 6 sets of writing scripts, followed by 6 sets of speaking performances. Each set included 4 scripts,3 GCSE scripts and one CEFR benchmark script. Each judge had a unique combination of scripts.

The judging allocation design was identical across languages and skills. It was created so as to maximise the number of times each script was seen across judges and the number of comparisons overall, while minimising the number of times each judge saw a particular script. Around 95% of possible comparisons were achieved across judges, while each judge saw each script a maximum of 2 times. The CEFR scripts were seen 15 times and GCSE scripts 8 times on average across judges. This means that each CEFR script participated in around 46 paired comparisons on average, while this was 26 times on average for the GCSE scripts. Literature suggests that over 20 paired comparisons per object should be sufficient for a reasonable level of scale reliability (e.g. Verhavert et al., 2019) (see further details on this in Data analysis section). An example of a pack design is presented in Appendix F.

Table 7 *Key features of the judging allocation design (identical for each component and language)*

| N judges | N sets per judge | N scripts per set | Avg N comparisons per script | N comparisons (% possible comps) |
|---|---|---|---|---|
| 13 | 6 | 4 | 26 (GCSE)<br>46 (CEFR) | 468 (94%) |

Electronic versions of scripts (including audio recordings) were assigned to sets and participants according to the judging allocation design. These were accessed electronically via a secure shared area to which the participants were given access, and rank ordering was conducted electronically. Each participant was assigned an electronic recording form for each component, which they completed with their ranks.

Typically, rank ordering exercises in other contexts are conducted on paper. Thus, there was a possibility that our participants would find it problematic to do this exercise electronically. During their familiarisation activities, which preceded the rank ordering exercise, the participants were given the opportunity to access practice scripts and rank order them electronically. At the end of their familiarisation they were asked whether they found rank ordering the 4 electronic files feasible. As the figure below demonstrates, an overwhelming majority of the participants in each language agreed or strongly agreed that the task was feasible. This provided some reassurance that undertaking a rank ordering task electronically, at least with only 4 scripts in each set to rank order, was unproblematic and does not invalidate the rank ordering outcomes.

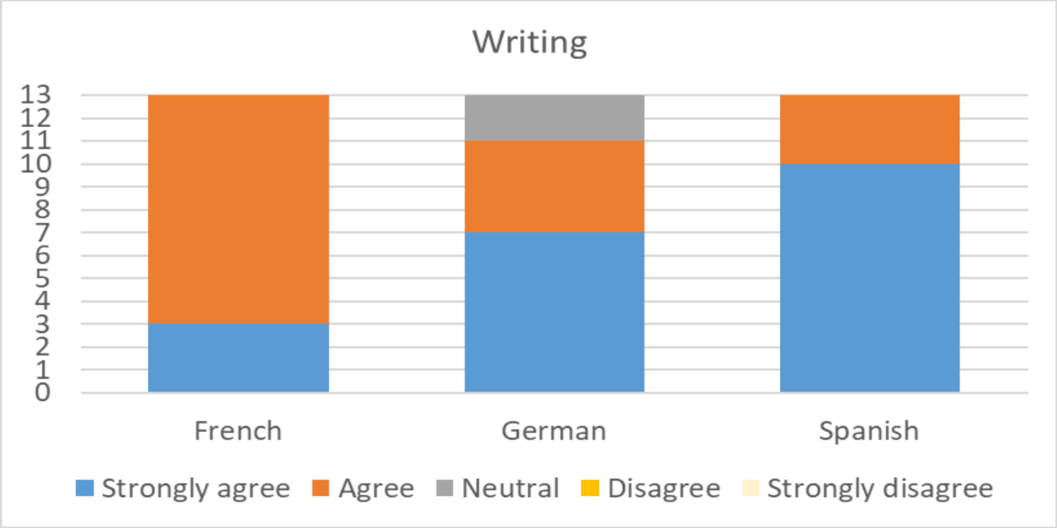


Figure 17 *“I found rank ordering 4electronic files (writing or speaking) feasible”*

The participants were provided with detailed instructions about how to complete the rank ordering task. An Ofqual researcher and IT support were also on hand in case of any issues with accessing the files, etc. Sample instructions and recording forms are included in Appendix G.

The participants were asked to complete familiarisation activities for productive skills before undertaking the rank ordering exercise. They were also asked to first complete rank ordering for writing before moving on to speaking. They were instructed to place the scripts in each set into a single rank order, from best (rank 1) to worst (rank 4), based on a holistic judgement of overall quality. It was emphasised that they were to consider the important aspects of language competence according to the CEFR as the basis for their judgements of quality, even if these may be different from what is currently considered important for GCSE MFL qualifications and assessments. They were asked to try as best as they could to focus on the qualities of performances and try to ignore the fact that these were based on different examinations (i.e. the CEFR exemplars came from a range of different international exams, rather than GCSEs).

The participants were instructed to only use GCSE mark schemes for reference and that they were not to attempt to re-mark any GCSE scripts. Rather, they were asked to make a holistic judgment of the quality of each script relative to other scripts in the set.

With respect to the translation task, which forms part of the writing paper, our content mapping exercise suggested that use of CEFR written mediation scales as the basis for evaluation of the quality of translation may not be appropriate, as the GCSE translation tasks seemed focused primarily on testing vocabulary and grammar rather than other aspects of translation. Therefore, we advised the participants that, during rank ordering, they treat the translation performances as primarily evidence of vocabulary and grammatical competence rather than genuine mediation, though we suggested that if they disagreed with this view, they were free to use the mediation scale if helpful and rank the relevant performances accordingly.

# Standard linking of reading and listening comprehension assessments

The third activity, which was also carried out by all participants, involved "standard linking" of the reading and listening 2018 assessments (rather than performances) to the CEFR during panel meetings. It is standard practice that the linking of assessments for receptive skills is done with reference to assessment demand rather than performance quality, due to these assessments typically involving selected response and short-answer questions. Following Brunfaut and Harding (2013), we used a modified version of a standard setting technique called the Basket method to implement this (cf. the Manual, ibid.).

## *Rationale for the method*

The Basket method was chosen amongst a large number of alternative methods (cf. Cizek and Bunch, 2007; the Manual, ibid.) because it was deemed both more realistic in terms of the level of judgement precision that the participants can make about test questions and less cognitively demanding. Most other standard setting methods require panellists to estimate chance levels or proportions of minimally competent candidates likely to answer a particular question correctly. These estimates are often criticised in the literature as unlikely to be correct. In our context, this would have been even more problematic in the absence of IRT-based item statistics which would have helped to inform the difficulty rank order of items, thus perhaps helping the participants to make more reliable judgements as in the Bookmark method. Furthermore, because only one test was available, from the first administration in 2018, we believed that it was more appropriate to consider this test as an exemplar of other possible tests based on the same specification, rather than an established test on which standards can be set in more explicitly quantitative ways, using robust statistical evidence.

The Basket method only requires the participants to rate each task in terms of the first CEFR level at which candidates could reasonably be expected to respond to tasks like that correctly, effectively placing each task, as an exemplar of similar tasks, into a CEFR level "basket". The modified Basket method additionally requires panellist to select a sub level for a task after deciding on the CEFR level first. The sub-levels are derived from the CEFR scales, with low, mid and high sub categories (see below for more details on this).

By summing and weighting these ratings we can derive a set of cut scores on the GCSE tests related to the CEFR scale. The approach taken for calculating the cut-scores followed Brunfaut and Harding (ibid.) and de Jong (2009) in using weightings that imply a "comfortable" achievement of a particular level. Setting a cut score at a comfortable level means that the candidates would be expected to answer 50% of the tasks rated at that level correctly, and 80% or more of the tasks rated at the levels below. In other words, a candidate achieving a particular level threshold would have 50% chance of answering the tasks at that level correctly and 80% or higher chance of answering the tasks at the level(s) below correctly. We will explain the details of the cut score calculation method further in the Data analysis section. This approach was taken in order to guard against a common criticism of the Basket method that it tends to result in cut scores that are too lenient (cf. the Manual, ibid., Kaftandjieva, 2010).

Once the CEFR related cut scores are set in this way, they can be compared to the 2018 GCSE grade boundary marks and distributions of candidates achieving these. This would allow us to estimate where current grade boundary marks and their associated performance standards are in relation to the cut scores derived based on the CEFR performance scales.

## Procedure

The standard linking activities were conducted in panel meetings over 2 days. A week prior to the meetings the participants completed the relevant familiarisation activities, as described previously. According to the advice from the Manual, the panel considered the reading comprehension test on the first day, followed by the listening test, which is often more difficult to judge reliably, on the second day. The panel sessions were audio recorded with the consent of the participants.

As part of the familiarisation, as well as at the start of the panel meetings, different aspects of the standard linking activities and rationale for those were explained to the participants, including:

- the key question of the Basket method
- the notion of the “task” as the focus of the ratings
- the focus on task demand rather than difficulty,
- how to rate partial credit tasks (i.e. tasks where 2 or more marks can be achieved), and
- the scale to be used for categorising tasks and mark points (where question tariff was 2 or more marks) and the meaning of the scale categories in relation to the CEFR

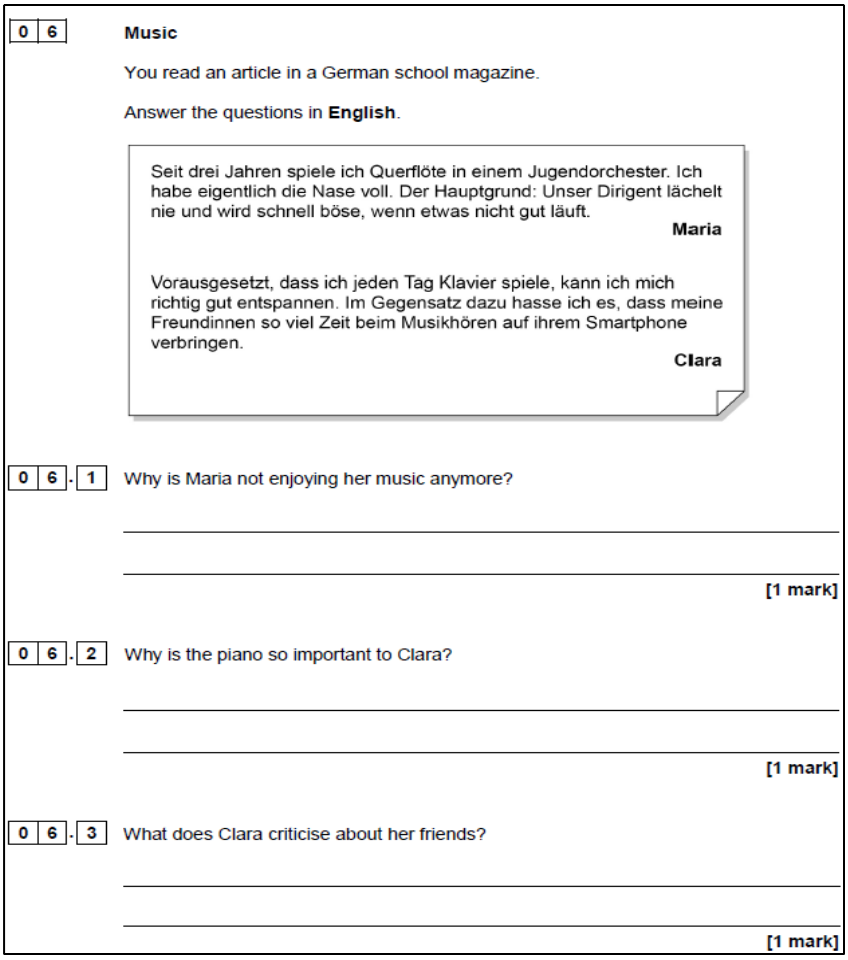
The key question that each participant needed to answer in order to categorise each task was:

***"Which is the first CEFR level describing learners who can answer this task correctly?”***

This question refers to the task rather than the exam question as it is usually not possible to consider the demand of the text and an associated question completely separately. It is usually necessary to consider the whole task, which involves the question, relevant parts of the text which the question refers to, any intentionally distracting parts of the text, and the interaction between these elements.

In addition, where it is possible to achieve partial credit on a task, it is necessary to consider the CEFR level of learners that would score each possible creditworthy mark point. In some cases, the learners at the same level or sub level would be able to achieve each mark, while in some cases only higher level learners would be able to achieve the higher marks.

As an example, in Figure 18, there are 3 tasks, i.e. 3 one-mark questions associated with a single text. Here, the participants gave a single rating for each task.



0 6 **Music**

You read an article in a German school magazine.

Answer the questions in **English**.

> Seit drei Jahren spiele ich Querflöte in einem Jugendorchester. Ich habe eigentlich die Nase voll. Der Hauptgrund: Unser Dirigent lächelt nie und wird schnell böse, wenn etwas nicht gut läuft.
>
> **Maria**
>
> Vorausgesetzt, dass ich jeden Tag Klavier spiele, kann ich mich richtig gut entspannen. Im Gegensatz dazu hasse ich es, dass meine Freundinnen so viel Zeit beim Musikhören auf ihrem Smartphone verbringen.
>
> **Clara**

0 6 . 1 Why is Maria not enjoying her music anymore?

**[1 mark]**

0 6 . 2 Why is the piano so important to Clara?

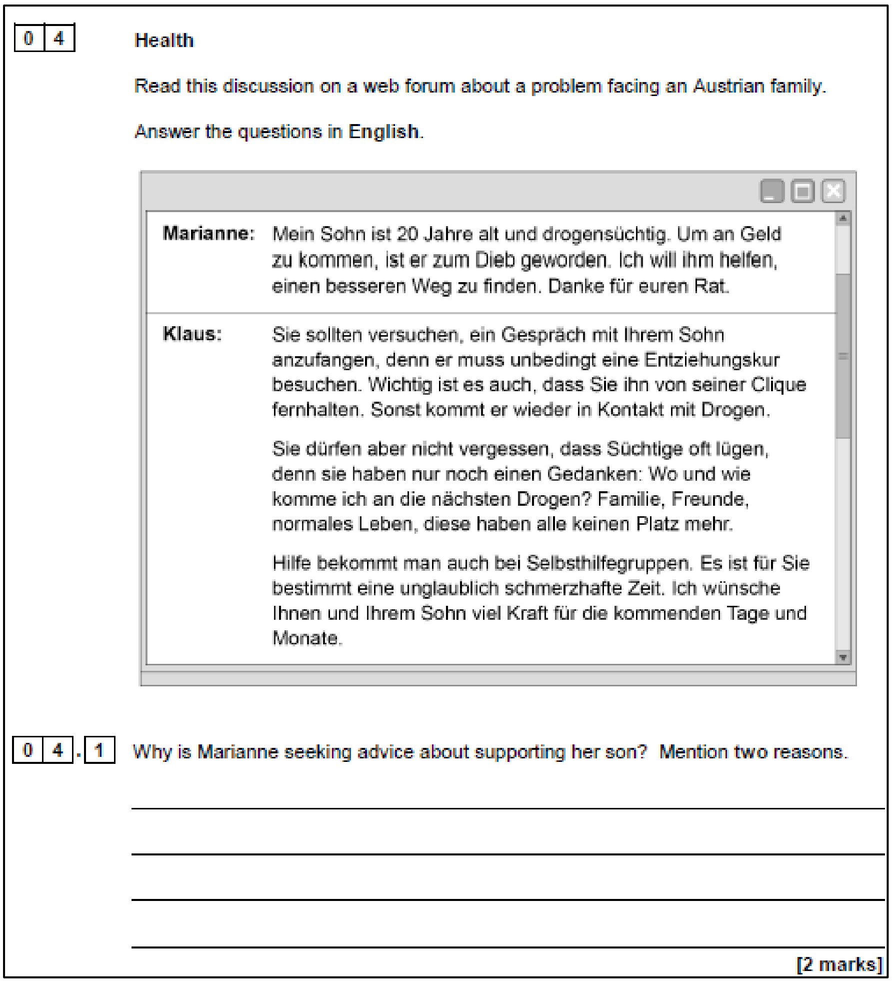
**[1 mark]**

0 6 . 3 What does Clara criticise about her friends?

**[1 mark]**

Figure 18 *Example of one-mark tasks*

In Figure 19, there is a task where it is possible to achieve one mark (partial credit) or 2 marks (full credit). Here, separate ratings were given for each creditworthy mark point. In other words, the panellists decided on the level of the learner who could get one mark and then on the level of the learner who could get 2 marks on this task. Thus, 2 separate ratings were given for this task.



0 4 **Health**

Read this discussion on a web forum about a problem facing an Austrian family.

Answer the questions in **English**.

**Marianne:** Mein Sohn ist 20 Jahre alt und drogensüchtig. Um an Geld zu kommen, ist er zum Dieb geworden. Ich will ihm helfen, einen besseren Weg zu finden. Danke für euren Rat.

**Klaus:** Sie sollten versuchen, ein Gespräch mit Ihrem Sohn anzufangen, denn er muss unbedingt eine Entziehungskur besuchen. Wichtig ist es auch, dass Sie ihn von seiner Clique fernhalten. Sonst kommt er wieder in Kontakt mit Drogen.

Sie dürfen aber nicht vergessen, dass Süchtige oft lügen, denn sie haben nur noch einen Gedanken: Wo und wie komme ich an die nächsten Drogen? Familie, Freunde, normales Leben, diese haben alle keinen Platz mehr.

Hilfe bekommt man auch bei Selbsthilfegruppen. Es ist für Sie bestimmt eine unglaublich schmerzhafte Zeit. Ich wünsche Ihnen und Ihrem Sohn viel Kraft für die kommenden Tage und Monate.

0 4 . 1 Why is Marianne seeking advice about supporting her son? Mention **two** reasons.

**[2 marks]**

Figure 19 *Example of a multi-mark task*

It was explained to the participants that, in order to answer the standard linking question for each task or mark point, they would need to form a judgement about the features that contribute to the demand of the task, and consider how these features would interact with the performance features of learners at a particular CEFR level and sub-level. It was highlighted that this judgement does not imply that learners at a lower level could not give the correct response; it means that (in the eyes of the panel member) a correct response should not reasonably be expected at lower CEFR levels.

It was emphasised that the focus of standard linking is the demand of each task based on its various features, effectively treating each task as an exemplar of other possible similar tasks that could appear in a test. This is because it is usually not possible to predict accurately how any task will be received by any specific sample of students in a specific test unless the questions are field tested appropriately. This is why it is usually conceptually simpler to focus on the key task features, apparent in the task itself, that are likely to affect demand rather than trying to estimate question difficulty, i.e., how this task may have performed on the actual test.

The participants were asked to rate tasks first in terms of broad CEFR levels, and then “fine-tune” the judgements in terms of sub-levels for each task. The levels and sub-levels used are presented in the table below.

Table 8 *CEFR levels and sub-levels used in standard linking*

| Level | Sub-level |
|---|---|
| Above B2 | |
| B2 | B2 high |
| | B2 mid |
| | B2 low |
| B1 | B1 high |
| | B1 mid |
| | B1 low |
| A2 | A2 high |
| | A2 mid |
| | A2 low |
| A1 | A1 high |
| | A1 mid |
| | A1 low |
| Below A1 | |

Regarding the scale categories, it was explained to the participants that, in general, global CEFR descriptors summarise the threshold (criterion) abilities of learners that belong to a level (North and Schneider, 1998). Learners that do not have at least some level of the abilities listed in the descriptor for a particular CEFR level will be at the level below. Therefore, each task may be targeted at a threshold level, or somewhat higher but still within the same overall level. The “low” sub-level within each level was defined as the threshold level.

The “high” sub-levels were defined as the “plus” levels, which are used in some specific CEFR scales. The plus levels describe learners at the top of the level. These learners will have a full range of the threshold skills and abilities and possibly some elements of the level above. The participants were familiar with these from the scales used in their familiarisation. The “mid” sub-level was defined as somewhere in between the threshold and plus levels.

The panellists all had their own copies of the question papers, mark schemes, and the relevant CEFR scales that were provided to them during familiarisation. The ratings were collected using the same survey tool which was used during familiarisation. An excerpt from the tool is presented in Appendix H.

After the general concepts and aspects of the standard linking procedure were explained and discussed, the panellists had the opportunity to discuss the collated ratings given to the CEFR exemplar tasks as well as the sample of GCSE tasks during familiarisation. This discussion provided an opportunity for the participants to raise any issues about the key rating question, the focus on demand vs. difficulty, the scale categories, etc. This discussion was led by one of the CEFR experts who carried out the content mapping exercise.

Following this, the first 3 sets of tasks in the question paper were considered jointly by the panel. This provided further opportunity for discussion and fine-tuning of their common understanding of the CEFR scales and sub-scale categories and how to relate these to task demands.

The remaining tasks were rated by each panellist independently. As is common in most standard-setting methods, these independent ratings constituted the initial round of judgements.

The initial ratings were then collated and the distribution charts of ratings for each task and mark point presented to the participants for discussion, prioritising those where there was most disagreement. The discussion was facilitated by Ofqual researchers, one per panel. During the discussion, the panellists were asked to justify their ratings for each task, and to consider the rationale of other panellists for theirs.

After the discussion, the panellists were asked to consider their ratings again independently and amend any ratings where they were convinced and could see that their own rationale may not have been appropriate, based on panel discussion. It was emphasised that they did not need to amend their ratings if they could not see a good reason to do so. In other words, there was no need to agree with the others despite their own personal views. This second round of judgements then constituted the final set of ratings.

At the end of the second day, the panellists were encouraged to reflect upon the CEFR linking exercise, consider the fit and misfit between the CEFR concepts of communicative language competence and language learning and GCSE, as well as the sources of demand in the current GCSEs. In addition, while the main focus of the panels was to conduct the standard-linking exercise, a great deal of discussion emerged spontaneously at several point during panel meetings in relation to the task at hand, but also in relation to the nature of GCSE MFL assessments and teaching practice, purpose of the MFL GCSEs etc. The main points from this discussion are summarised in the Qualitative results section.

# Data analysis

## *Rank ordering*

The rank orders obtained from the judges in the rank ordering exercise were converted into paired comparisons and a single ‘perceived quality’ scale across the GCSE and CEFR benchmark scripts was derived using a Rasch formulation of Thurstone’s (1927) paired comparison model (Andrich 1978; see Bramley, 2007).[16]

[16] Using this model to analyse the data can lead to over-estimation of the statistical separation of the objects on the latent trait because the ranking constrains the possible paired comparisons outcomes, leading to violation of the assumption of local independence in the model (cf. Bramley, 2005). This violation should not affect the resulting rank order of scripts significantly though. There is currently no reliable estimate of the amount of over-estimation this causes. The findings from a small replication study (Curcin and Black, in prep.), where the same judges conducted 2 rank ordering and paired comparison exercises on the same set of scripts, and with similar number of comparisons per script in each exercise, suggests that the scale of over-estimation might not be so large as to invalidate rank ordering outcomes. The rank ordering sets contained 6 scripts each. The separation coefficients obtained from rank ordering were 5.94 and 6.07. The corresponding paired comparisons coefficients were 4.69 and 4.20 respectively. The SSRs for the 2 rank ordering exercises were 0.97 and 0.97 for each paper respectively, compared to 0.96 and 0.95 in the paired comparisons exercises. Furthermore, the correlations between the measures from the rank ordering and paired comparisons exercises were 0.93 and 0.95, suggesting that there was little change in the rank order of the same scripts obtained from these exercises. It might be reasonable to expect that the amount of over-estimation would be less in the current study, given that there were only 4 scripts in each set, and that the rank order was not significantly affected by the violation of local independence.

Each script is positioned on this scale in terms of quality, which is related to the probability of it being judged better than another script in a paired comparison. The model can be stated as:

$$\ln[P_{ij} / (1 - P_{ij})] = B_i - B_j$$

where $P_{ij}$ = the probability that script *i* beats script *j* in a paired comparison
and $B_i$ = the measure for script *i*
and $B_j$ = the measure for script *j*

The unit of the script quality scale is known as a ‘logit’ or ‘log-odds unit’. The analysis was carried out using the Facets software version 3.66.3 (Linacre 2010). After the initial run, some data cleaning was undertaken where appropriate, in the following ways:[17]

- most misfitting observations were removed from analyses (based on highest standardised residuals for individual paired comparisons). In order to preserve most of the data, we tended to remove misfitting observations rather than all judgements from a judge or script that showed some misfit. However, judgements of 4 judges across 3 different papers were removed entirely to improve overall fit
- all scripts which won or lost all their comparisons, and hence had imputed measures, were removed from the plots of mark on measure and mark-measure correlation analyses

The results based on the cleaned data were evaluated in terms of model fit, scale properties and mark-measure correlation, as is standard in rank ordering exercises, and as explained below (based on Curcin and Black, ibid.).

A key way of establishing whether a rank ordering exercise has worked is to check the properties of the scale of perceived quality created by the judges. This involves investigating scale separation reliability (SSR) and model fit, which are the usual checks conducted for any latent trait analysis (cf. Bond and Fox, 2007).

The SSR coefficient is analogous to the person separation reliability in Rasch modelling (e.g. Andrich, 1982) and to KR-20, Cronbach Alpha, and the Generalizability Coefficient. It is calculated as:

$$\text{SSR} = \frac{(\text{Observed SD})^2 - \text{MSE}}{(\text{Observed SD})^2}$$

where Observed SD is the standard deviation of the estimated measures, and MSE is the mean squared standard error of the estimated measures across all the scripts.[18]

In this context, SSR means "reproducibility of relative measure location" (cf. Winsteps Manual https://www.winsteps.com/winman/reliability.htm). In our context,

[17] Details of data cleaning are presented in Appendix I.

[18] Separation coefficient is the ratio of the person true SD (i.e., the "true" standard deviation), to RMSE, the error standard deviation. It provides a ratio measure of separation in RMSE units, which is easier to interpret than the reliability correlation, with no upper bound as with SSR. Separation coefficient is the ratio of "true" variance to error variance. The relationship between separation coefficient and SSR is: separation coefficient = square-root(SSR/(1-SSR)) (cf. https://www.winsteps.com/winman/reliability.htm).

high reliability of the script measure scale would mean that there is a high probability that those scripts estimated with high measures actually do have higher measures (i.e. better quality) than the scripts estimated with low measures.

In general, the decision of whether the SSR of a scale can be considered satisfactory will depend on the purpose for which the scale is constructed, as well as on the context and type of the assessment under consideration. Verhavert et al. (2019) cite 0.7 as the level mentioned in the literature as appropriate for low-stakes or formative assessments, and 0.9 as the level often accepted as appropriate for high-stakes and summative assessments (Nunnally, 1978). In the rank ordering studies carried out to date, SSR of around 0.8 and higher has generally been judged as satisfactory and related to other aspects of the comparative judgement exercises being judged as satisfactory too.

A common way of checking overall model fit is to check the overall proportion of misfitting judgements. Usually, this should be at or below what would be expected by chance, i.e. less than about 5% of standardised residuals using the criterion of 2 for the absolute value of the standardised residual, and less than about 1% using the criterion of 3 (cf. Linacre, 2011). In addition to that, it is usually necessary to check the usual Rasch fit statistics for the scripts and judges (e.g. Wright and Linacre, 1994).[19] In particular, reasonable fit statistics of the judges would suggest the consistency of their judgements and a reasonable level of agreement on rank orders across all the judges.

In addition to these, checking the mark-measure correlation is a way to establish whether the judges in a comparative judgement exercise were perceiving a trait that is sufficiently similar to the one underlying the test scores. Previous rank ordering studies tended to consider correlations around and above 0.7 as satisfactory.

Once the appropriate script quality measures are obtained, their logit scale can be plotted against GCSE mark scale and the CEFR benchmark scripts and GCSE scripts placed on the logit scale to observe their relative position and extrapolate the likely performance standards of the relevant GCSE grade boundary scripts in relation to the CEFR benchmark performance standards.

## Standard linking

Following Brunfaut and Harding (ibid.), the sub-level scale with low, mid and high levels was transformed into continuous numerical scale shown in Table 9.

Table 9 *Numerical rating scale categories – CEFR sub-levels*

| Sub-level | Numerical score |
|---|---|
| Below A1 | 0 |
| A1 low | 0.67 |
| A1 mid | 1 |
| A1 high | 1.33 |
| A2 low | 1.67 |

[19] Note the limitations of Rasch-based fit statistics with respect to unknown exact sampling distributions (e.g., Christensen, et al., 2013; Karabatsos, 2000; Smith, Schumacker and Bush, 1998). However, useful applications of these indices have been demonstrated in the literature (e.g. Wright and Linacre, 1994), and their use for exploratory or descriptive purposes may be considered appropriate despite the limitations (e.g., Engelhard, Kobrin and Wind, 2014).

| | |
|---|---|
| A2 mid | 2 |
| A2 high | 2.33 |
| B1 low | 2.67 |
| B1 mid | 3 |
| B1 high | 3.33 |
| B2 low | 3.67 |
| B2 mid | 4 |
| B2 high | 4.33 |
| Above B2 | 5 |

Once the sub-level ratings were transformed into the numerical scale, mean sub-level ratings for each task and score point were calculated. These mean ratings were then transformed back into the CEFR level ratings, using the ranges shown in Table 10. The rationale for these ranges was that, as each CEFR level was conceptualised as containing low, mid and high compartments, the best way of classifying ratings at each compartment would be to divide the level into 3 equal parts and select the mid-point of each of these parts as the scale point.

Table 10 *Numerical rating scale categories - CEFR levels*

| Level | Score range |
|---|---|
| A1 | 0.51-1.50 |
| A2 | 1.51-2.50 |
| B1 | 2.51-3.50 |
| B2 | 3.51-4.50 |

The cut scores were then set based on sums of weighted frequencies of the CEFR level ratings. The weightings were based on the notion of each cut score representing the level at which a candidate can answer 50% of the tasks at that level correctly; 80% of the tasks at the level below; and 95% of the tasks at the levels below that. This is based on typical Item Response Theory probabilities, which was the method by which CEFR descriptors were scaled in development (cf. Brunfaut and Harding, ibid.; de Jong, ibid.).[20]

For instance, based on the following frequencies of tasks rated at each CEFR level, the cut score for A2 level would be calculated as 1x0.80+37x0.50=19.3 (rounded to 20). The rounding applied was to the next larger integer.

Table 11 *Example frequency table from which cut scores are calculated*

| CEFR level | N marks | Cut Score |
|---|---|---|
| A1 | 1 | 1 |
| A2 | 37 | 20 |
| B1 | 22 | 42 |

The analysis of the results of the standard linking task involved checking that the outcomes are based on reasonably reliable and consistent judgements. This was established using intra-class correlations (ICCs) (cf. e.g. Hallgren, 2012). All ICCs were two-way because there was no resampling of raters for each item, and based on average scores. We report both agreement-based ICCs (i.e. showing agreement

[20] Another approach would be to calculate the cut score for a “just-qualified” candidate, which would be established from a count of the number of items below a particular level + 1. In the above example, this would mean that A2 cut score would be 2, and B1 cut score 38.

level in absolute terms), and consistency-based (i.e. showing agreement in terms of rank order of values).

We first calculated the ICCs for initial ratings (excluding the items which were used for familiarisation and initial discussion in the panels). This was to estimate the level of agreement before discussion took place and ratings were changed as a result of discussion. ICCs were then also calculated for the final ratings, on which the cut score analysis was based.

## *Qualitative analysis*

The audio recordings of the standard-linking panels were transcribed. Analysis of the transcripts was carried out by 2 researchers using a framework approach to identify salient themes from individual panels as well as common themes between the 3 panels. The analysis involved the following steps:

1. Familiarisation: Both researchers read through a transcript sample to gain familiarity with the data. During this process both researchers made notes on any emergent and obvious themes from the transcript for later discussion.
2. Identifying a thematic framework: Both researchers met to discuss notes about the transcript and agree on a thematic framework for coding of remaining transcripts.
3. Indexing: Researchers then coded the remaining transcripts using the framework. The framework was not definitive and as such additional concepts and themes emerging from ongoing coding were noted, discussed and investigated in all the transcripts for decision on their inclusion.
4. Charting: Researchers met again to compare additional notes and discuss preliminary findings. During this phase, the researchers highlighted the themes common between the panels and grouped codes into related categories.
5. Mapping and interpretation: With the help of 3 additional researchers, who were involved in the moderation of the panels, themes and codes were discussed to ensure analysis was relevant to the study aims. Additional searches were also conducted following these meetings to investigate further emergent themes.

# Limitations

Because this study was designed to answer a specific research question rather than as a full-blown linking study, it consequently has some potential limitations in scope and generalisability. This is, to our knowledge, the first explicit, albeit limited in scope, attempt to link GCSE MFL qualifications to the CEFR using recommended methodology, and so we consider this study primarily exploratory. Involvement of other relevant stakeholders (e.g. exam boards, Department for Education, etc.), greater resources, further refinement of some aspects of the methodology, and linking of specifications from other exam boards would be necessary to conduct an official linking study, the results of which might be considered to represent an "official" linking, endorsed by the key stakeholders. Therefore, the findings need to be treated as essentially descriptive and indicative.

Having said this, we have made every effort to conduct this linking study according to best practice in the field, and in this sense, the results should be reasonably robust for those specifications on which the linking was performed. However, below we highlight some of the specific methodological limitations that limit the generalisability and validity of our results and interpretations either across the board or for some specifications or components.

The linking was conducted on a sample of GCSE MFL specifications, one per language. Therefore, even though the content and grading standards across these and other available specifications should be aligned and are monitored through Ofqual's standard procedures, it is possible that the results of the linking for, for instance, AQA German may not perfectly generalise to another GCSE specification in German.

Another limitation is that the linking was conducted using only one, as well as the first, instance of the relevant examinations from newly reformed specifications. Therefore, the performances of students in this examination, as well as the nature of the receptive skills tests, may have reflected the novelty of the curriculum and assessments and may not be fully representative of a "steady state". On the other hand, Stratton and Zanini's (2019) research, in which they compared the functioning of the 2017 unreformed and 2018 reformed GCSE MFL assessments, suggests that some of the novel features did not always appear to have tangible effects. Furthermore, Cuff's research (2018) into the sawtooth effect when new specifications are introduced, suggests that these effects are fairly small. Nevertheless, for a full linking study to be conducted appropriately, it would be ideal for experts to have access to more than one instance of the examination paper, alongside the relevant specification. If an official linking study was to be conducted, it would be advisable to do this at the point when the new curriculum and assessment methodology are reasonably embedded.

It is possible that the results might have benefitted from more training and familiarisation activities for productive skills. However, we do believe that, based on the statistical indicators of judge consistency and scale reliability, the results of this part of the linking study did not particularly suffer from this limitation and can, for the most part, still be considered reasonably trustworthy.

Another issue specifically relevant for the rank ordering study, and mostly affecting the results for French, is the possibility that the CEFR benchmark performances used in the ranking exercise may not have been fully representative of the relevant CEFR levels for French. We did not know this in advance of the exercise, but the consultant from Institut Français, who carried out the content mapping, subsequently suggested that the CEFR exemplars available on the Council of Europe website may be slightly dated in some respects and suggested that current exemplars from Institut Français be used in future linking exercises. Unfortunately, this situation makes it to some extent difficult to interpret the linking results for French productive skills. This is further discussed in the relevant results section.

Especially with reference to the procedure for carrying out the linking for receptive skills, as with any other situation where human judgement is elicited from a group of experts, it is possible that various group effects may have been at play and somewhat affected the overall results. There is a possibility that aspects of group dynamics such as conformity (e.g. Asch, 1951; Deutsch and Gerard, 1955), polarisation (i.e., adoption of a more extreme position) (Moscovici and Zavalloni, 1969), and, to an extent, “group think” (Baron, 2005) may have created more of a consensus than might have been the case in a different set up. On the other hand, the purpose of those discussions was indeed to try and achieve consensus regarding the relevant standards as far as this was appropriate given judges’ individual views. In a high stakes linking study, or a standard setting exercise, it would be ideal to utilise 2different panels of experts for each specification and explore the extent to which their independent results agreed and supported each other. In our study, the fact that the results for comparable qualifications, albeit from different languages, by and large replicate each other, despite being arrived at based on the judgements of independent panels, could be taken as indication that the panellists’ judgements and our procedures were sound and derived reasonably robust results.

With respect to the qualitative analysis and findings, the panel discussions were focussed on linking receptive skills assessments to the CEFR and reflecting on the findings of the linking exercises. Consequently, their main focus was not to discuss broader issues systematically. This may have focussed discussion to very specific areas and therefore not allowed much exploration of other topics which may also be relevant in the context of MFL learning and assessment.

As the sample of experts volunteered to participate in this project, and were selected specifically because of some familiarity with the CEFR, they may be relatively more informed or intrigued by the topics discussed in comparison to randomly selected experts. Therefore, the views of these panels may be different to their peers. This does not, however, mean they are any less valid or relevant.

The generalisability of our qualitative analysis may be limited due to the broadly inductive, theory generating approach taken. Although the mapping phase of analysis allowed some testing of the emergent themes, this was limited. Further robustness to our conclusions could be gained through future research aimed at investigating the themes identified further, and considering any further emergent themes.

# Results

## Content mapping

The experts that conducted content mapping using the CEFR scales and categories did so assuming the communicative, action-oriented approach to language learning and assessment. This is the approach advocated in the CEFR and reflected in the aspects of language competence that are described through its various descriptor scales. As already noted in the introduction, it was deemed appropriate to consider the GCSE MFL assessments in relation to the CEFR, given that the approach suggested in the curriculum for GCSE MFLs is intended to be communicative. However, the content mapping exercise highlighted some discrepancies in the way some aspects of language competence are assessed in GCSE compared to what might be expected of assessments intended to assess communicative language competence, whether or not they are aligned to the CEFR.

Some of these inconsistencies were taken into account when the CEFR level ratings were given to different elements of assessments. For instance, in the listening comprehension test for German, even though the breadth of contexts might have warranted level B1, the way the assessment was operationalised, with very short texts and a significant amount of scaffolding in the questions, limited the ability of the test to assess detailed understanding, interpretation and inference. This effectively lowered the level of the skills that could be demonstrated to high A2, with some elements of B1.

The tables below present the results of content mapping for the 3 languages to the CEFR. The slashes indicate the categories where the specification, the contexts, or the texts in the comprehension tests suggest a higher level of demand, i.e. B1, but the actual operationalisation in the 2018 tests, the nature of questions and tasks, the nature of assessment criteria, or use of English in responses effectively reduces the level that can be demonstrated (e.g., A2).

In general, the consultants noted that there was little in the way of integrated skills assessed explicitly. Therefore, these are not presented separately in the tables. As for mediation skills, the only explicitly assessed mediation activity in the test is translation. However, the way that it is assessed, i.e. mostly word for word and essentially designed to assess knowledge of grammar and vocabulary, does not allow explicit use of CEFR mediation scales to rate it. Therefore, this task was rated with respect to linguistic competence scales.

Across the board, the experts noted that there are very few aspects of the specification or tasks in the test that could be accessed by learners at A1 level. Overall rating for French and German was given as up to A2+/low B1, and for Spanish as low B1.

What is particularly notable in the tables below is that the consultants were not able to rate the productive skills assessments for written interaction for French and German because there were no tasks that explicitly assessed that aspect. For example, no written requests have to be made and there are no messages, letters, notes or forms to be completed in the exam, nor were they mentioned in the specification. In Spanish, the consultants did give a rating for written interaction, though they noted that the scope for assessing this skill seemed limited in this specification and paper.

Spoken interaction was given a rating but it was noted that, given the very structured nature of the spoken interaction tasks and the fact that the teacher initiates the conversations/communication, there were limited opportunities for genuine interaction to fill information gaps. For this reason, candidates cannot demonstrate some of the abilities required for B1, notably “Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest”. The exchanges mainly involve answering questions. The fact that the candidate can only ask one or two questions limits the possibility of showing interactive skills and thus limits the level of ability that can be demonstrated.

It was further noted that the amount of prepared spoken production did not allow candidates to use their full linguistic range while being spontaneous. The test format seems to elicit shorter answers rather than longer turns. Only a minority of descriptors for sustained monologue (describing experiences and putting a case) relate to the test format. The subscales for sustained monologue also seem to point to A2 rather than B1. There was evidence that candidates were rewarded for using memorised phrases and expressions instead of communicating by assembling their own syntactic structures and using their own words. It was also noted that the format of the spoken test does not allow for wider scope of sociolinguistic proficiency, such as acting according to local politeness conventions of which candidates would be aware.

With respect to receptive skills, it was noted that the overall levels of the texts used tended to be higher than the level of the questions. In Spanish, some texts were characterised as B2 level. In reading, the tasks tended to focus on reading for information. Most listening tasks were characterised as quite short, making it difficult to separate detailed understanding from understanding key messages (gist). Tasks aiming at detailed understanding were described as very structured. Where texts were more abstract, and of higher level, the questions are asked in English. This was seen as not coherent with the communicative approach and it effectively lowers the CEFR level of the task. The experts noted that sociolinguistic and pragmatic skills were not explicitly tested, but could be assumed to be there as they are required to understand the context of the text for instance.

Speaking rate in the listening test was characterised as not really standard speech for German and Spanish, but a range of topics justifies low B1 to some extent. In Spanish, the consultants also noted that the slow speed of speech did not always lower the difficulty of the question, as the difficulty was in the topic and language used.

Table 12 *Content mapping ratings for productive skills*

| | Overall | Overall spoken interaction | Overall spoken production | Overall written interaction | Overall written production | Linguistic | Socio-linguistic | Pragmatic | Strategic | Mediation |
|---|---|---|---|---|---|---|---|---|---|---|
| French | A2+/B1 low | A2+/B1 low | A2+ | n/a | A2+/B1 low | A2+/B1 low | A2+/B1 low | A2+/B1 low | A2+/B1 low | A2 |
| German | A2 | A2 | A2/B1 | n/a | B1 | B1 | A2 | A2 | A2 | A2 |
| Spanish | B1 | B1 | B1 | B1 | B1 | B1 | B1 | B1 | B1 | B1 |

Table 13 *Content mapping ratings for receptive skills*

| | Overall | Overall listening comp | Overall reading comp | Processing Text | Linguistic | Socio-linguistic | Pragmatic | Strategic | Mediation |
|---|---|---|---|---|---|---|---|---|---|
| French | A2+/B1 low | A2+/B1 low | A2+/B1 low | A2+/B1 low | A2+/B1 low | A2 | A2 | A2 | A2 |
| German | B1 | A2/B1 | A2/B1 | B1 | B1 | B1 | B1 | B1 | A2 |
| Spanish | B1 | B1 | B1 | B1 | B1 | B1 | B1 | B1 | B1 |

To summarise, the following were the discrepancies or gaps noted compared to what might be expected in assessments intended for communicative language competence and fully aligning with the CEFR notion of communicative language competence:

1. Absence of explicit assessment of interactive skills in writing (except to some extent in Spanish). This would be relevant for assessment at both A2 and B1 level of competence.
2. Little potential for genuine interaction to fill information gaps in speaking given the scripted nature of the assessments and overly prescriptive mark schemes; relevant for assessment at both A2 and B1 level of competence.
3. No explicit assessment of integrated skills, already relevant for assessment at A2 level.
4. With respect to German and Spanish listening comprehension in particular, the rate of speech was characterised as not standard, which would be a requirement for a test that truly assessed up to level B1.
5. The issue of questions in the receptive skills tests being in English was highlighted as contrary to the communicative approach, and effectively reducing the potential of these assessments to assess in the way that would provide evidence of candidates' communicative competence especially at B1 level. In addition, the experts noted that this approach may not necessarily be helpful for candidates as it turns some of the tasks into translation exercises, even though it may be easier for test developers to develop questions in English.
6. It was noted that the translation task was not really congruent with the CEFR mediation scales as translation is just one strand of mediation. Our French consultant noted, however, that mediation is not normally assessed fully at A2 level, and is more appropriate for assessment starting from B1 level.

Overall, the experts thought that there was enough construct overlap for a content mapping and linking exercise to be appropriate despite the discrepancies in the nature of assessments noted above. However, they emphasised that it was necessary to acknowledge these discrepancies, especially where, even though content might warrant assessment at higher levels, assessment operationalisation effectively lowered the scope and level of the skills that could be demonstrated by candidates, and hence the interpretation of the resulting scores.

This suggests that there might be scope for current GCSE assessments to make the way they assess certain communicative language abilities more effective and congruent with the communicative approach. This is particularly prominent with respect to a lack of appropriate assessment of spoken and written interaction, as well as integrated skills, which could be seen as representing core skills to be acquired if the MFL curriculum and pedagogy are indeed to "develop [learners'] ability to communicate confidently and coherently with native speakers in speech and writing, conveying what they want to say with increasing accuracy" (DfE, ibid.: 3).

Broadly speaking, the experts agreed that each of the 3 GCSE MFL specifications considered assessed most of the skills up to A2+ (i.e. high A2) level, with some aspects of language competence assessed up to low B1 level. However, it should be borne in mind that the limitations of assessments noted above, particularly with respect to assessment of interaction, would to some extent limit the interpretation based on these assessments that candidates are fully at A2 or B1 level. In particular, most of the caveats and discrepancies noted above relate to where assessments appear to be targeting B1 level, as in many cases assessments were patchy in the extent to which they allowed for all of the skills relevant for B1 level to be demonstrated.

# Rank ordering of written and spoken performances to map to the CEFR

## *Evaluation of Rasch model fit and scale properties*

In order to have confidence in the results of a rank ordering exercise, it is necessary to evaluate the model fit and script quality scale properties in particular (see Data analysis section for more details on this).

Overall model fit can be seen in Table 14 for each language and skill. In general, model fit was satisfactory, with less than 5% standardised residuals greater than absolute 2 and around 1% standardised residuals greater than absolute 3.

Table 14 *Overall model fit*

| | | StRes > abs 2 | | StRes > abs 3 | |
|---|---|---|---|---|---|
| Language and skill | N valid observations | N | % | N | % |
| Spanish writing | 926 | 16 | 1.73 | 6 | 0.65 |
| Spanish speaking | 860 | 32 | 3.72 | 16 | 1.86 |
| German writing | 786 | 16 | 2.04 | 10 | 1.27 |
| German speaking | 926 | 38 | 4.10 | 10 | 1.08 |
| French writing | 842 | 26 | 3.09 | 4 | 0.48 |
| French speaking | 898 | 12 | 1.34 | 6 | 0.67 |

Individual judge fit was also satisfactory, with most infit mean squares between 0.5 and 1.5, suggesting that the judges were reasonably consistent in their judgements.[21] Script fit statistics were also largely reasonable, consistent with the overall satisfactory model fit. [22]

The scale separation reliability (SSR) and separation were high in each case, suggesting that the observed differences between scripts were not due to measurement error. This is shown in Table 15.

Table 15 *SSR and separation coefficients*

| | Writing | | Speaking | |
|---|---|---|---|---|
| Language | SSR | Separation | SSR | Separation |
| Spanish | 0.96 | 5.21 | 0.91 | 3.23 |
| German | 0.92 | 3.99 | 0.91 | 3.10 |
| French | 0.95 | 4.22 | 0.96 | 4.91 |

There was generally good agreement between the GCSE mark and measure scale across all grades, as well as between the CEFR benchmark script rank order and their quality measures. This is shown in Table 16. The French CEFR rank-measure

[21] Judge fit statistics are presented in Appendix J.
[22] Script statistics are presented in Appendix K.

correlations were somewhat lower than for the other subjects, resulting from the measures for B1 and B2 benchmarks for writing, and A2 and B1 benchmarks for speaking being reversed (i.e., the higher CEFR level benchmark was perceived to be of lower quality than the lower level benchmark). The consensual ratings from our panel collected post hoc suggested that the 2 reversed writing benchmarks were perceived to both be at B1 level, with the B1 benchmark perceived as mid B1, and the B2 one as high B1 level. As for the speaking benchmarks, according to official ratings, the A2 benchmark was described as A2+ (i.e. high A2). Our panel consensual ratings judged it as mid A2, while they judged the B1 script as low B1. This could to some extent account for the reversal in the rank ordering exercise as the scripts may have been perceived to be closer in quality than the official ratings suggest.

Table 16 *Mark/rank order-measure correlations*

| | Writing | | Speaking | |
|---|---|---|---|---|
| Language | Mark-measure | CEFR rank-measure | Mark-measure | CEFR rank-measure |
| Spanish | 0.92 | 0.94 | 0.82 | 0.96 |
| German | 0.89 | 0.98 | 0.84 | 0.98 |
| French | 0.79 | 0.87 | 0.80 | 0.93 |

Some GCSE or CEFR scripts ‘won’ or ‘lost’ all of their comparisons and thus had imputed measure with large standard errors. These scripts did not contribute to measure estimation of other scripts. It was deemed appropriate to remove these scripts from further analyses as their imputed measures could not be deemed to represent a realistic measure of quality.

## *Mapping GCSE performances to the CEFR*

### *Writing*

Table 17 shows the mark points of the GCSE scripts included in the rank ordering exercise for writing. Recall that there were 3 scripts on each of the mark points below in the ranking exercise (i.e., 3 scripts with mark 18, 3 with mark 17, etc.). These mark points are also presented on the charts below.

Table 17 *Mark points of writing scripts included in the rank ordering exercise*

| Language | Grade 4 and below | Grade 7 and below | Grade 9 and below |
|---|---|---|---|
| Spanish | 18, 17, 16 | 34, 33, 32 | 49, 48, 47 |
| German | 22, 21, 20 | 40, 39, 38 | 50, 49, 48 |
| French | 26, 25, 24 | 44, 43, 42 | 53, 52, 51 |

The figures below show the relationship between individual GCSE script quality measures (x axis) and the measures of CEFR exemplars, denoted by orange dots. The CEFR ratings in brackets denote consensual panel average rating.[23]

[23] The charts for each language represent independent scales, from 3 independent rank ordering exercises, with their own samples of CEFR benchmark scripts, and were not linked by means of common scripts between different languages. This is why the x axes for each language are slightly different and cannot be compared with the others in absolute terms. The key comparison to make is

It can be seen that, while there are clear increases in perceived quality with increasing script grade, there is some variability in script quality between scripts around each individual grade boundary relative to their original mark. At each grade boundary, some grade boundary scripts were perceived to be of lower quality than some scripts on lower marks. There was also some overlap in perceived quality between scripts on different grade boundaries. For French writing in particular, while there is a clear increase in perceived quality between grades 4 and 7, there is very little differentiation in perceived quality between scripts at grades 7 and 9.

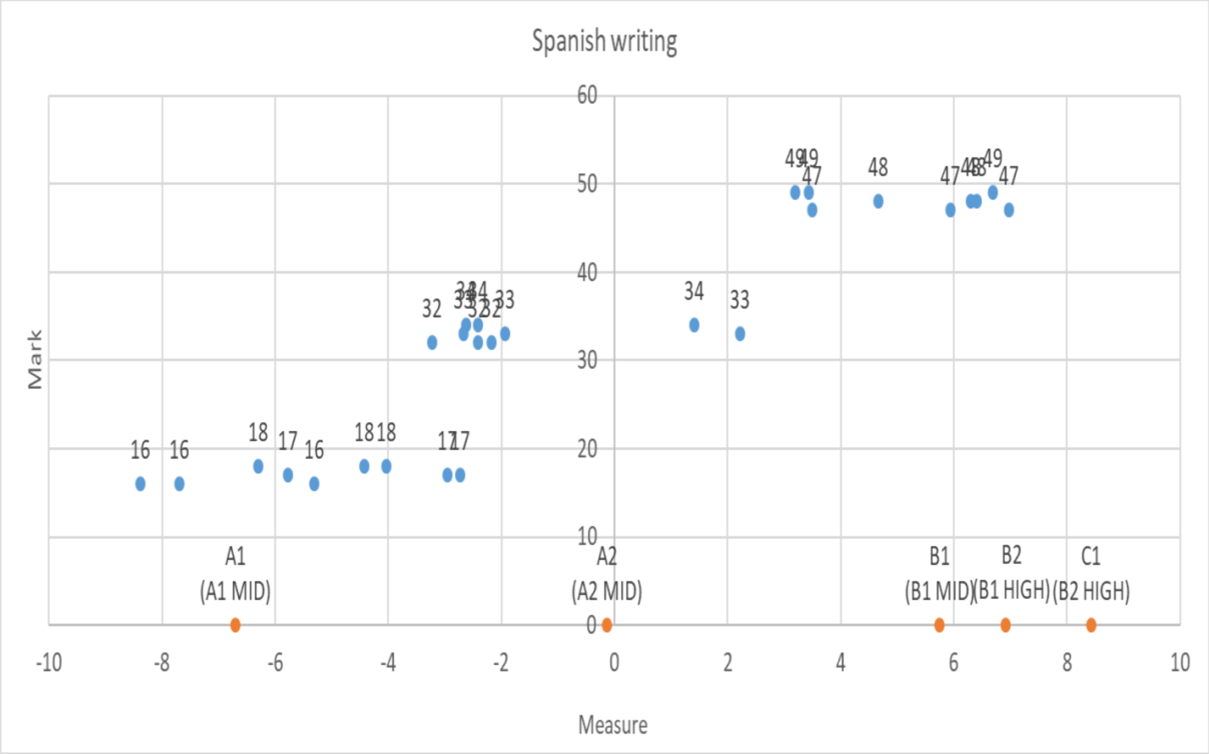


Figure 20 *Spanish writing rank order - individual script measures*

with respect to relative position of the GCSE and CEFR scripts on each scale. This applies to both writing and speaking.

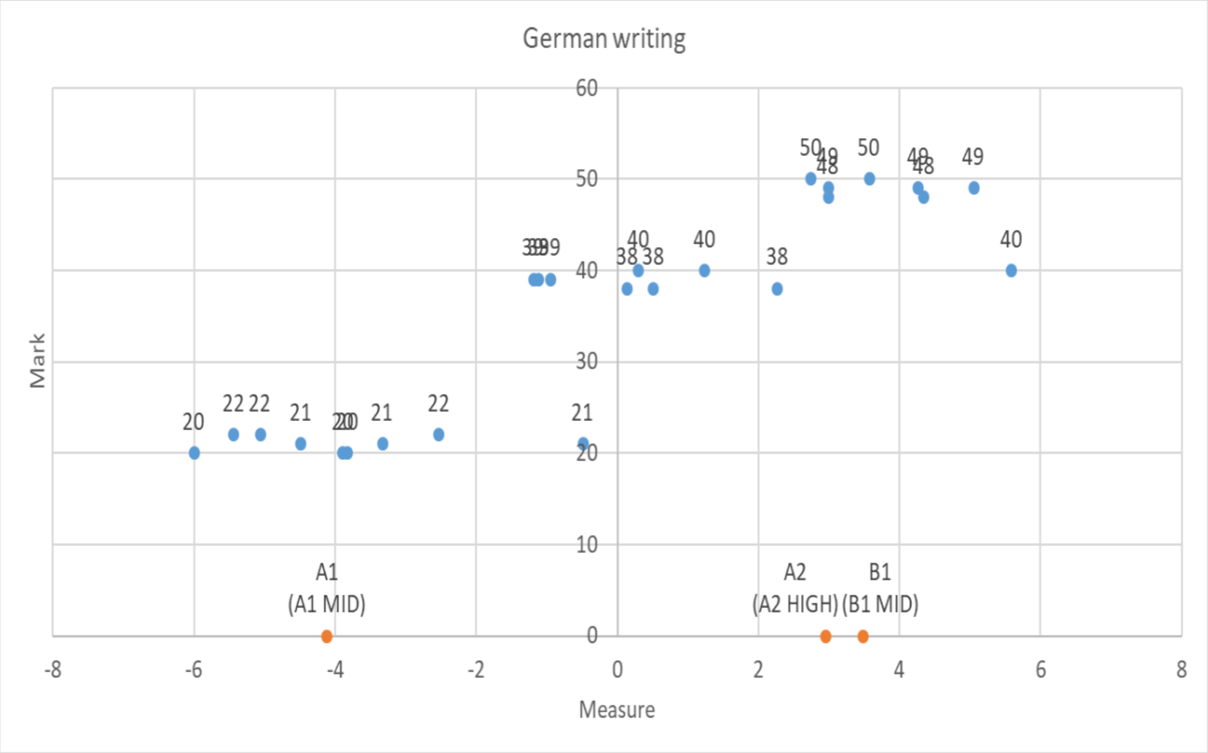


Figure 21 *German writing rank order - individual script measures*

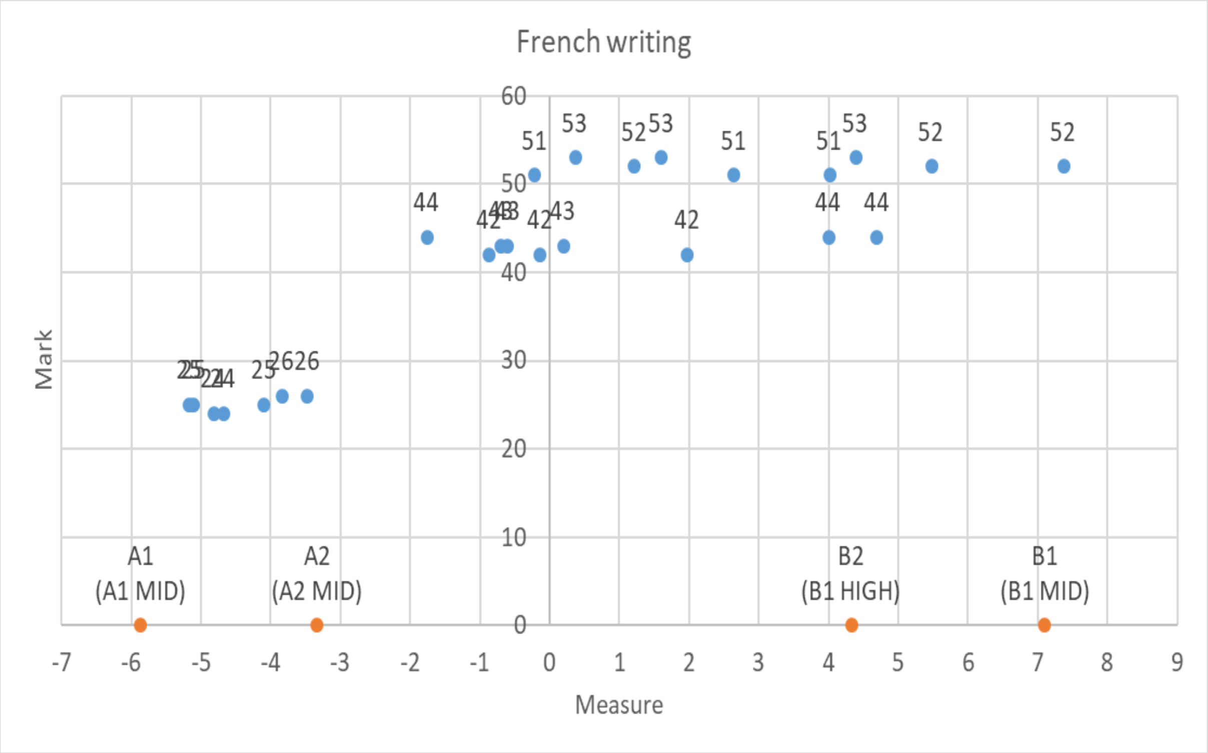


Figure 22 *French writing rank order - individual script measures*

Despite the above-mentioned script-level variability, it is possible to get a sense of where on average each grade boundary script set (i.e. across the three scripts on each grade boundary) is situated relative to the CEFR benchmark scripts. This is shown in the charts below for each language.

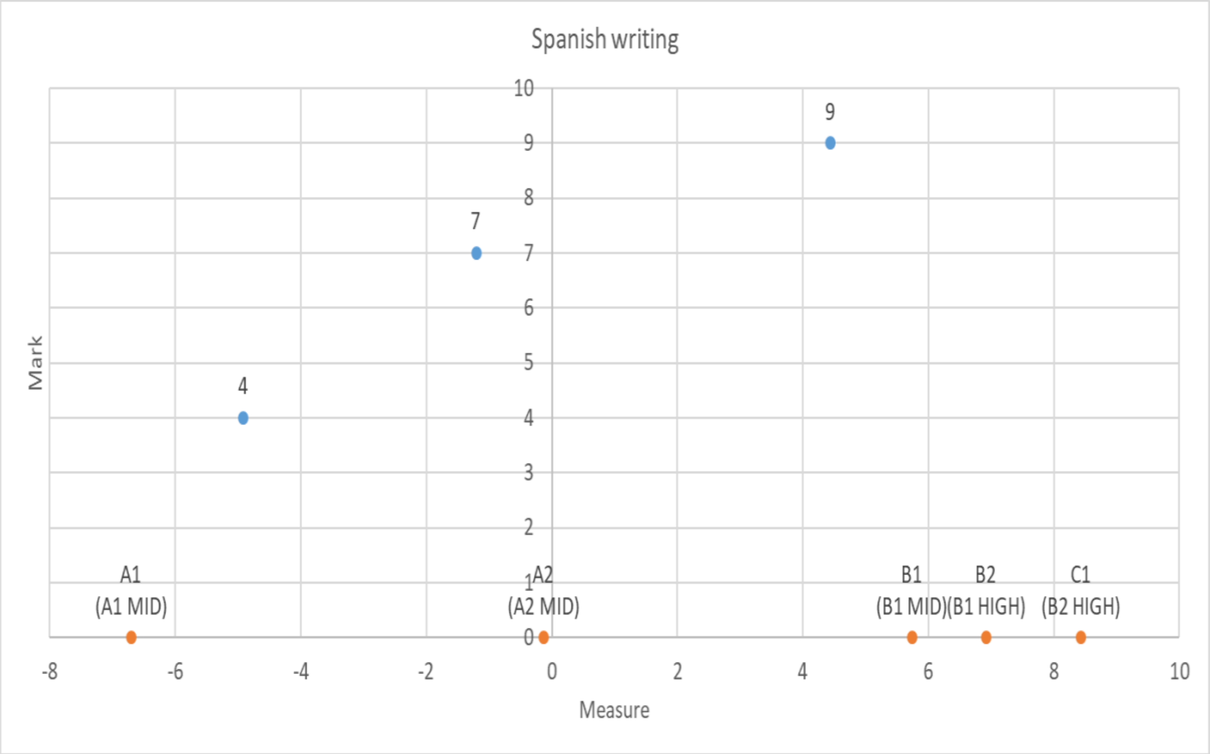


Figure 23 *Spanish writing rank order - average grade boundary script measures*

In Spanish, in relation to official CEFR ratings, it appears that grade 4 scripts are close in quality to the A1 benchmark. The grade 7 scripts are close to A2 benchmark. The grade 9 scripts are close to B1 benchmark. Our panel benchmark ratings suggest that officially A1 benchmarked CEFR script can be described as mid A1, which would suggest that the grade 4 scripts might be more precisely described as at mid to high A1 level. Along the same lines, the grade 7 scripts are low to mid A2, and grade 9 scripts low to mid B1.

Therefore, the GCSE grades can be mapped approximately to the CEFR as follows:

Table 18 *GCSE to CEFR mapping for Spanish writing*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | Mid-high A1 | A1 |
| 7 | Low-mid A2 | A2 |
| 9 | Low-mid B1 | B1 |

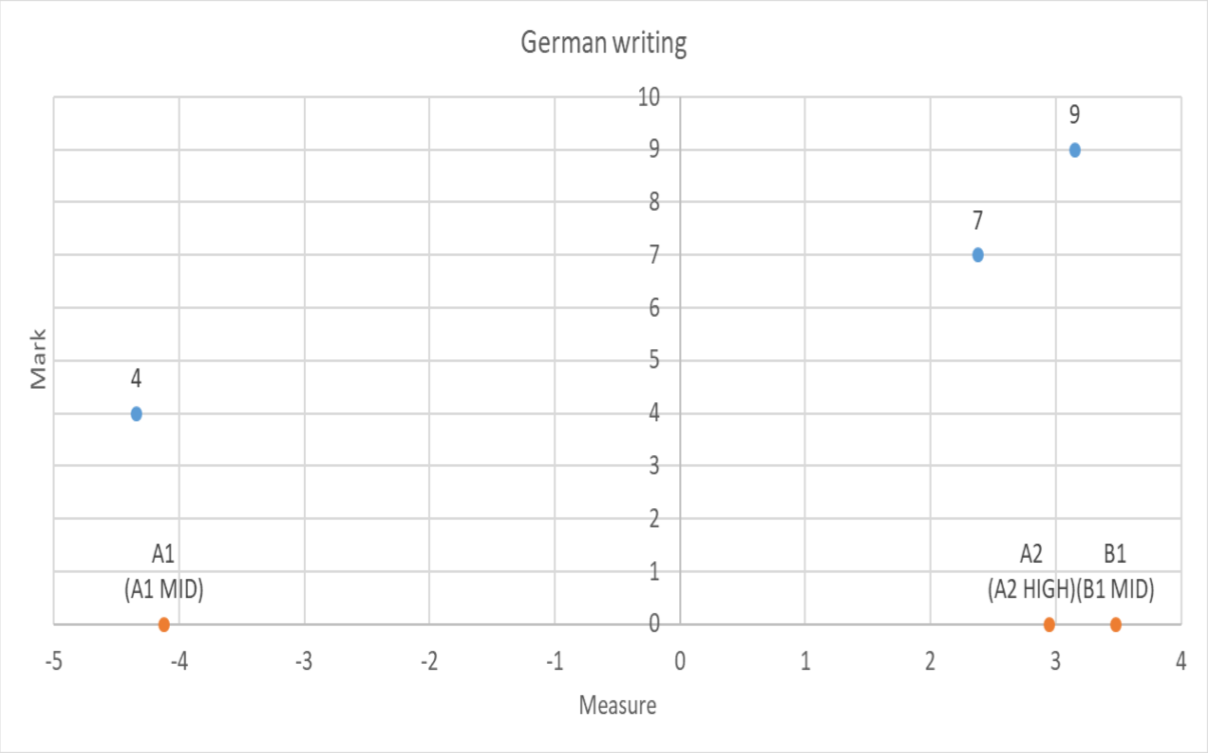


Figure 24 *German writing rank order - average grade boundary script measures*

In German, in relation to official CEFR ratings, it appears that grade 4 scripts are close in quality to A1 benchmark. The grade 7 scripts are close to A2 benchmark. The grade 9 scripts are close to B1 benchmark. Our panel benchmark ratings suggest that the officially A1 benchmarked CEFR script can be described as mid A1, which would suggest that the grade 4 scripts might be more precisely described as low to mid A1 level. Along the same lines, the grade 7 scripts are mid to high A2, and grade 9 scripts low to mid B1.

Therefore, the GCSE grades can be mapped approximately to the CEFR as follows:

Table 19 *GCSE to CEFR mapping for German writing*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | Low-mid A1 | A1 |
| 7 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 |

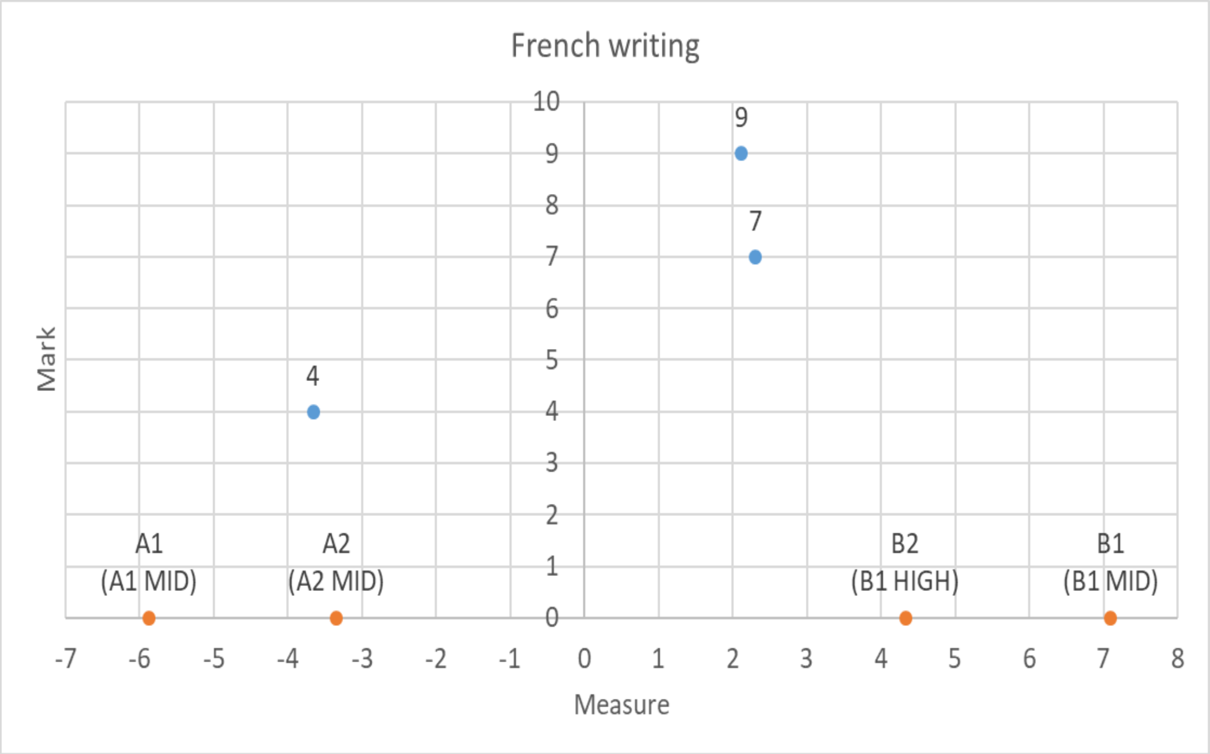


Figure 25 *French writing rank order - average grade boundary script measures*

In French, in relation to official CEFR ratings, it appears that grade 4 scripts are close in quality to A2 benchmark. The grade 7 and grade 9 scripts are almost identical in perceived quality. Unfortunately, given the issues about reversal of B1 and B2 benchmarks, it is difficult to judge which CEFR level they are closest to. But given the logit distances obtained for the other 2 languages between A2 and B1 levels, grades 7 and 9 might be somewhere in the region between B1 low and B2 high benchmark. Our panel benchmark ratings suggest that the officially A2 benchmarked CEFR script can be described as mid A2, which would suggest that the grade 4 scripts might be more precisely described as high A1 to low A2 level. Along the same lines, the grade 7 and grade 9 scripts might be low to mid B1.

Therefore, the GCSE grades can be mapped approximately to the CEFR as follows:

Table 20 *GCSE to CEFR mapping for French writing*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | High A1-Low A2 | A1-2 |
| 7 | Low-mid B1 | B1 |
| 9 | Low-mid B1 | B1 |

## *Speaking*

Table 21 shows the mark points of the GCSE scripts included in the rank ordering exercise for writing. Recall that there were three scripts on each of the mark points below in the ranking exercise. These mark points are also presented on the charts below.

Table 21 *Mark points of speaking scripts included in the rank ordering exercise*

| Language | Grade 4 and below | Grade 7 and below | Grade 9 and below |
|---|---|---|---|
| Spanish | 20, 19, 18 | 43, 42, 41 | 62, 61, 60 |
| German | 27, 26, 25 | 43, 42, 41 | 52, 51, 50 |
| French | 27, 26, 25 | 44, 43, 42 | 53, 52, 51 |

As in writing, the figures below show the relationship between individual GCSE script measures (x axis) and the measures of CEFR exemplars, denoted by orange dots. The CEFR ratings in brackets denote consensual panel average ratings.

Similarly to writing, while there are clear increases in perceived quality with increasing script grade, there is some variability in script quality between scripts around each individual grade relative to their original mark. At each grade boundary, some grade boundary scripts were perceived to be of lower quality than some scripts on lower marks. There was also some overlap in perceived quality between scripts on and below grade 4 and grade 7 as well as between scripts on or below grade 7 and grade 9. For French, there was a clearer differentiation between grade 7 and 9 in speaking.

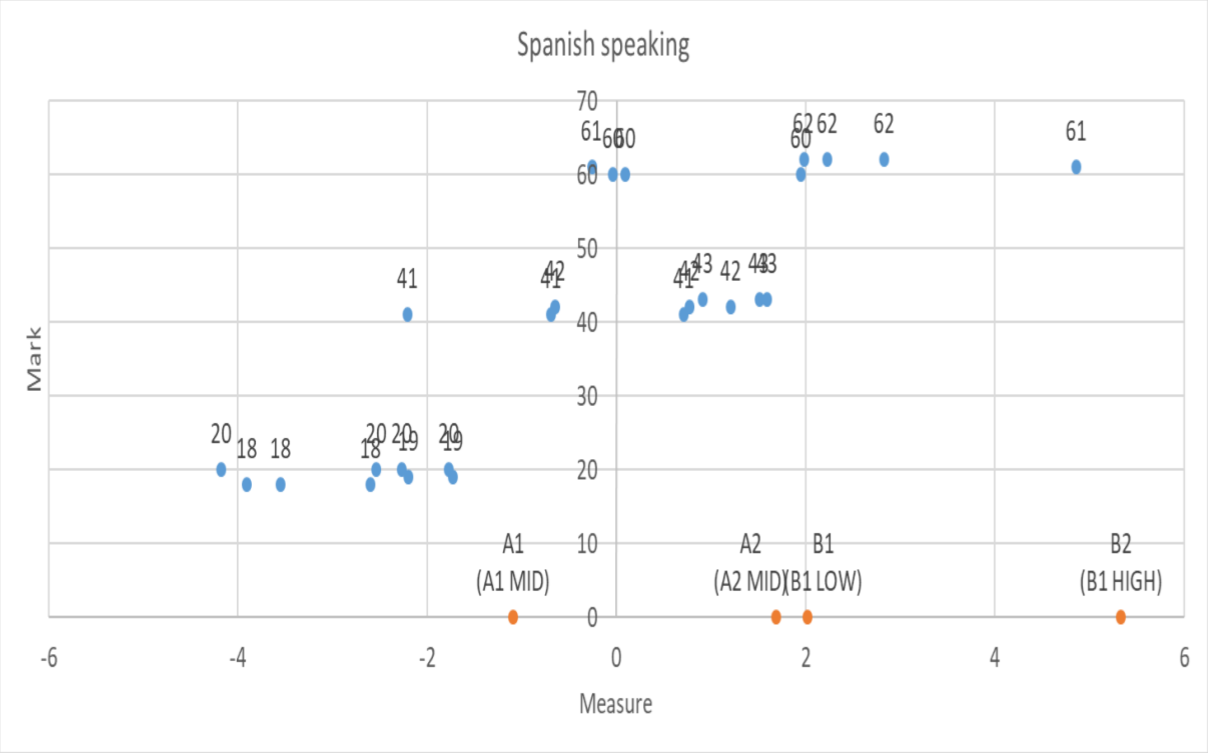


Figure 26 *Spanish speaking rank order - individual script measures*

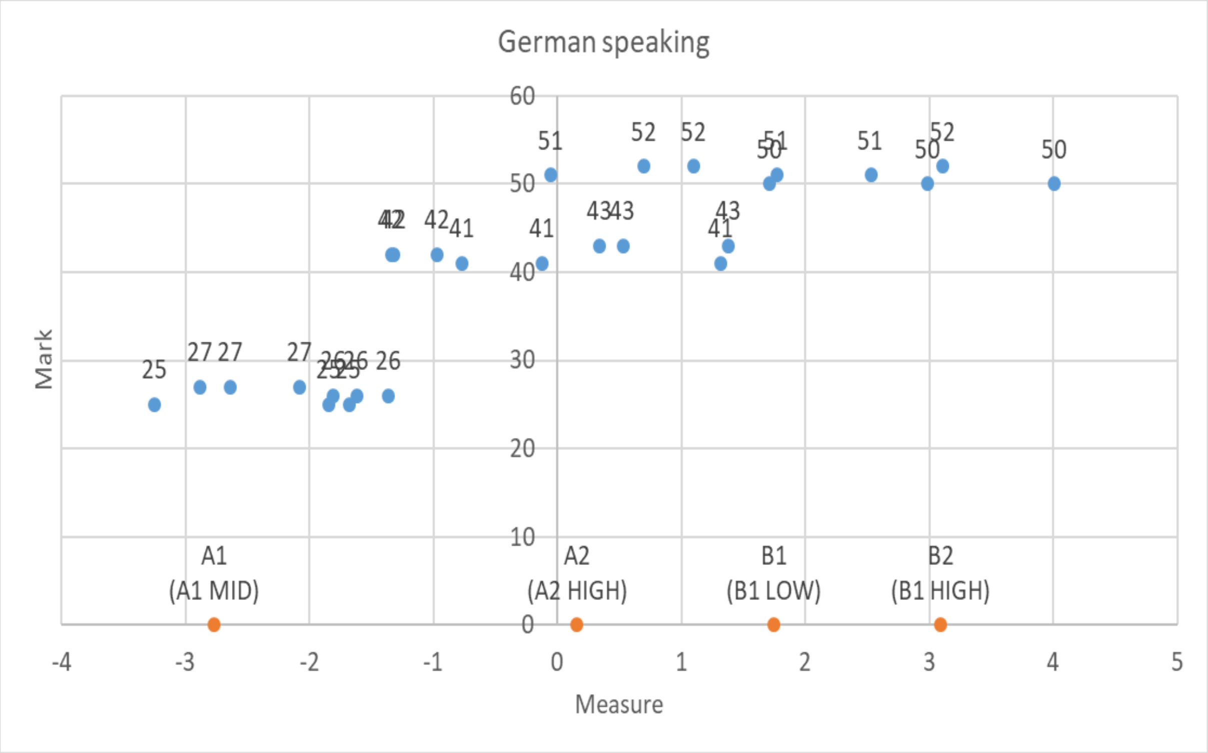


Figure 27 *German speaking rank order - individual script measures*

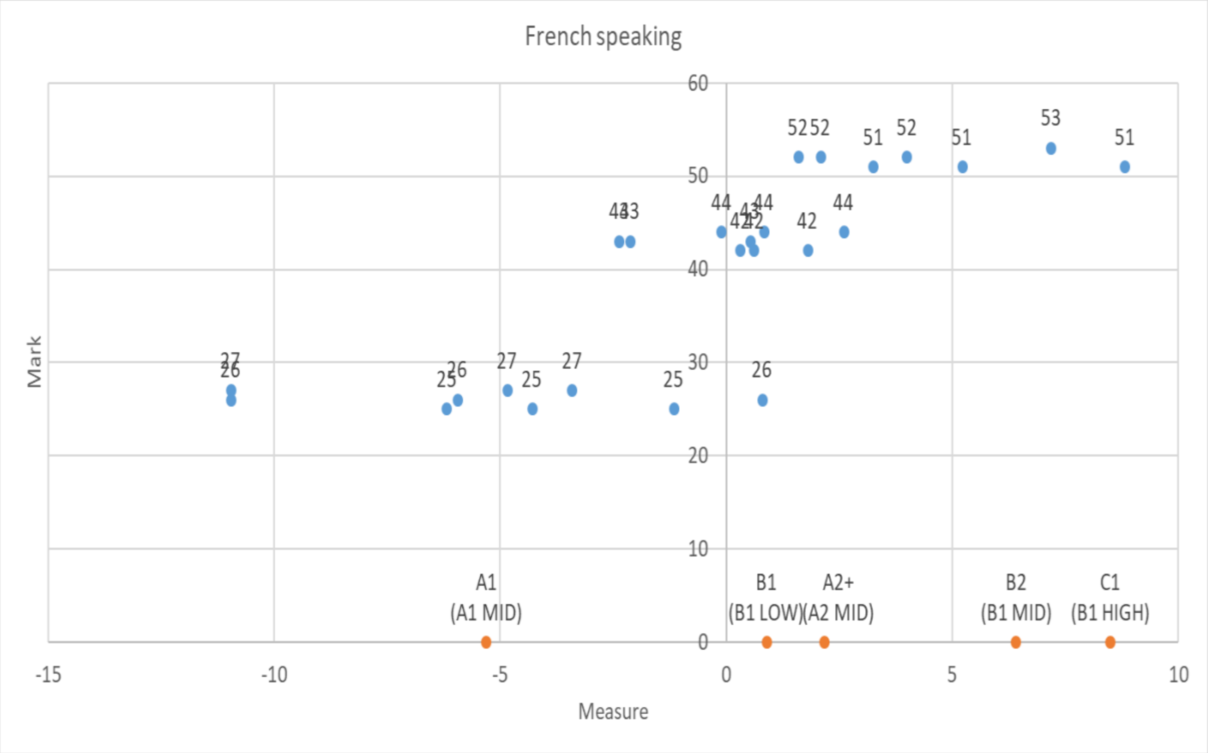


Figure 28 *French speaking rank order - individual script measures*

Despite the above-mentioned script-level variability, it is possible to get a sense of where on average each grade boundary script set (i.e. across the three scripts on each grade boundary) is situated relative to the CEFR benchmark scripts. This can be seen in the charts below for each language.

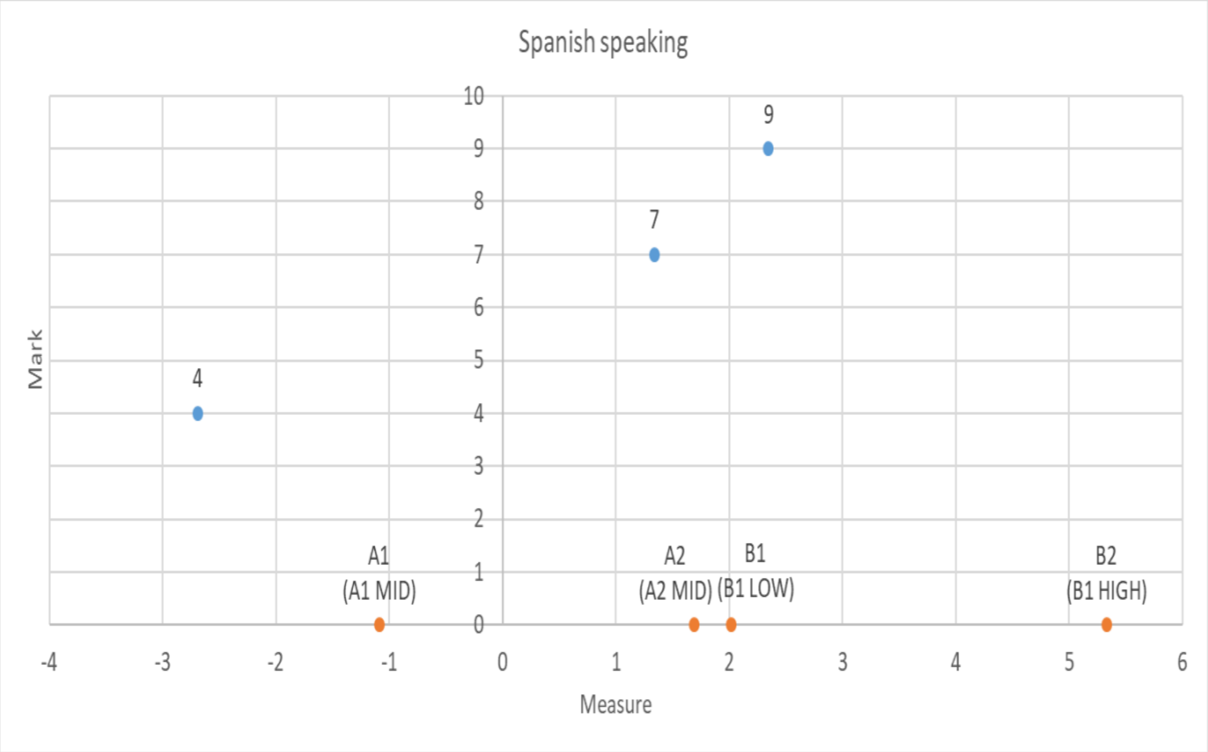


Figure 29 *Spanish speaking rank order - average grade boundary script measures*

In Spanish, in relation to official CEFR ratings, it appears that grade 4 scripts are quite a bit lower in quality from the A1 benchmark. The grade 7 scripts are close to A2 benchmark. The grade 9 scripts are close to B1 benchmark. Our panel benchmark ratings suggest that the officially A1 benchmarked CEFR script can be described as mid A1, which would suggest that the grade 4 scripts might be more precisely described as low to mid A1 level. Along the same lines, the grade 7 scripts are low to mid A2, and grade 9 scripts low to mid B1. Thus, we estimate that GCSE grade to CEFR level mapping for writing might be as follows:

Table 22 *GCSE to CEFR mapping for Spanish speaking*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | Low-mid A1 | A1 |
| 7 | Low-mid A2 | A2 |
| 9 | Low-mid B1 | B1 |

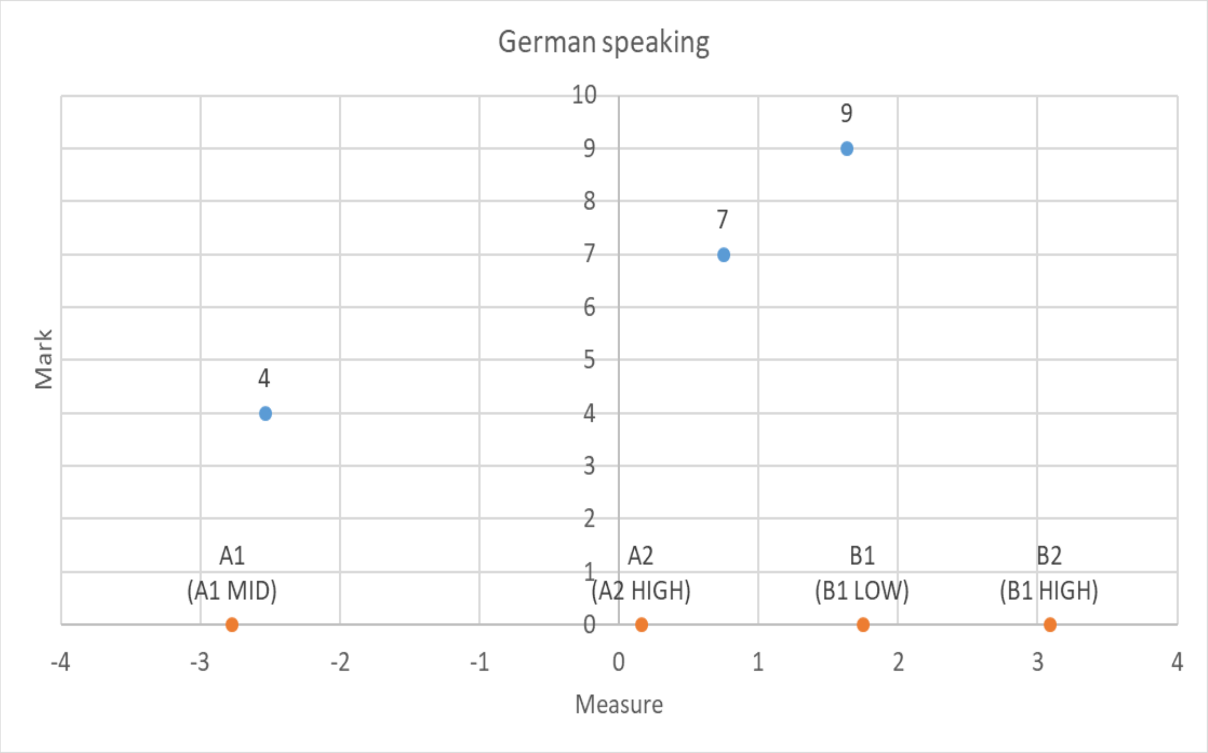


Figure 30 *German speaking rank order - average grade boundary script measures*

In German, in relation to official CEFR ratings, it appears that grade 4 scripts are close in quality to A1 benchmark. The grade 7 scripts are close to A2 benchmark. The grade 9 scripts are close to B1 benchmark. Our panel benchmark ratings suggest that the officially A1 benchmarked CEFR script can be described as mid A1, which would suggest that the grade 4 scripts might be more precisely described as mid to high A1 level. Along the same lines, the grade 7 and grade 9 scripts are high A2 to low B1.Thus, we estimate that GCSE grade to CEFR level mapping for writing might be as follows:

Table 23 *GCSE to CEFR mapping for German speaking*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | Mid-high A1 | A1 |
| 7 | High A2-low B1 | A2/B1 |
| 9 | High A2-low B1 | A2/B1 |

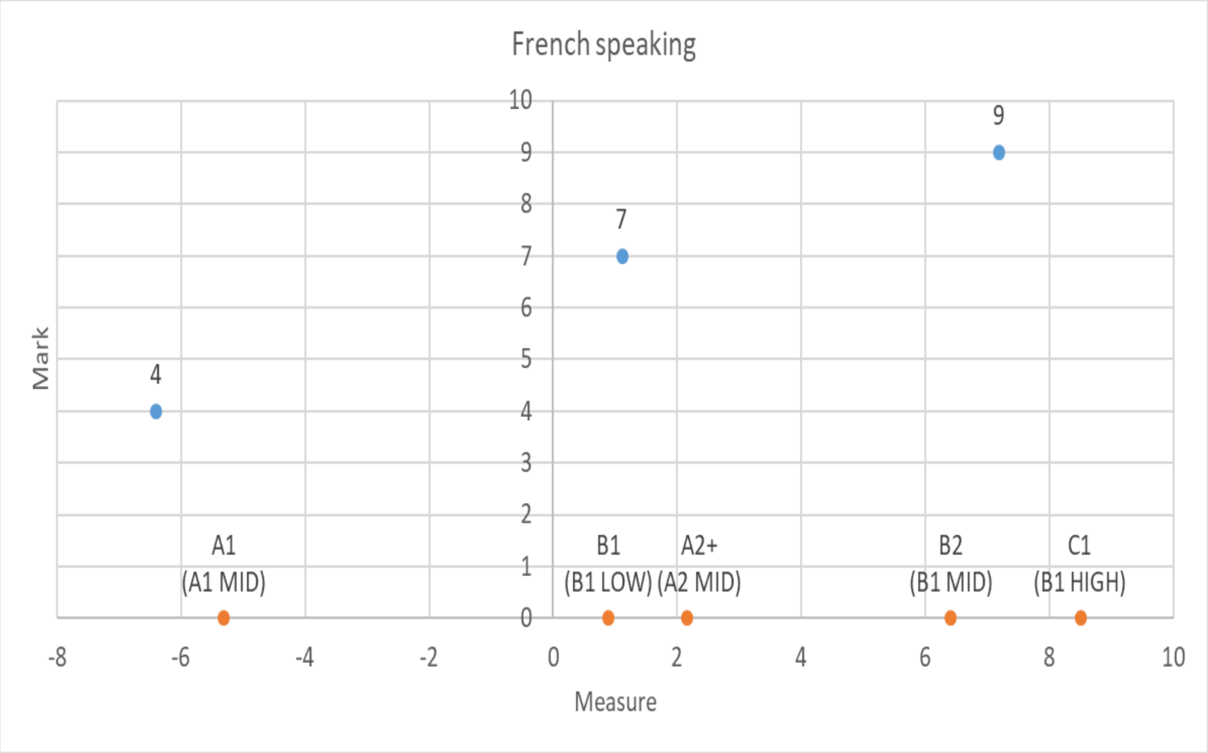


Figure 31 *French speaking rank order - average grade boundary script measures*

In French, in relation to official CEFR ratings, it appears that grade 4 scripts are close in quality to A1 benchmark. The grade 7 scripts are close to A2+/B1 benchmark. The grade 9 scripts are close to B2 benchmark. Our panel benchmark ratings suggest that the officially A1 benchmarked CEFR script can be described as mid A1, which would suggest that the grade 4 scripts might be more precisely described as low to mid A1 level. Along the same lines, the grade 7 scripts are high A2 to low B1 level. Interestingly, our panel considered the B2 benchmark of mid B1 quality and C1 benchmark of high B1 quality. Given the potential issue about the official ratings of the French benchmarks being obsolete, it may be appropriate to consider our panel ratings as more correct. This would make the grade 9 script mid to high B1. Thus, we estimate that GCSE grade to CEFR level mapping for writing might be as follows:

Table 24 *GCSE to CEFR mapping for French speaking*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | Low-mid A1 | A1 |
| 7 | High A2-low B1 | A2/B1 |
| 9 | Mid-high B1 | B1 |

## Summary and interim discussion

The tables below summarise the GCSE to CEFR mapping for productive skills.

Table 25 *GCSE to CEFR mapping for Spanish productive skills*

| | Writing | | Speaking | |
|---|---|---|---|---|
| GCSE grade | CEFR sub level | CEFR level | CEFR sub level | CEFR level |
| 4 | Mid-high A1 | A1 | Low-mid A1 | A1 |
| 7 | Low-mid A2 | A2 | Low-mid A2 | A2 |
| 9 | Low-mid B1 | B1 | Low-mid B1 | B1 |

Table 26 *GCSE to CEFR mapping for German productive skills*

| | Writing | | Speaking | |
|---|---|---|---|---|
| GCSE grade | CEFR sub level | CEFR level | CEFR sub level | CEFR level |
| 4 | Low-mid A1 | A1 | Mid-high A1 | A1 |
| 7 | Mid-high A2 | A2 | High A2-low B1 | A2/B1 |
| 9 | Low-mid B1 | B1 | High A2-low B1 | A2/B1 |

Table 27 *GCSE to CEFR mapping for French productive skills*

| | Writing | | Speaking | |
|---|---|---|---|---|
| GCSE grade | CEFR sub level | CEFR level | CEFR sub level | CEFR level |
| 4 | High A1-Low A2 | A1/2 | Low-mid A1 | A1 |
| 7 | Low-mid B1 | B1 | High A2-low B1 | A2/B1 |
| 9 | Low-mid B1 | B1 | Mid-high B1 | B1 |

Within Spanish, the writing and speaking assessments are quite consistent in terms of the mapping of GCSE grades to the CEFR levels. There were more discrepancies between writing and speaking within German and French. Straightforward direct comparisons between the 3 languages cannot be made as the scales of quality derived from the rank ordering exercise were not linked by design. However, on the assumption that the content and examination standards, as well as test specifications, are supposed to be reasonably aligned in these 3 languages, it might be informative to consider the results of the linking across the 3 languages.

It appears that the performance standards with respect to what students can do in CEFR descriptor terms, at grades 4 and 9 are reasonably aligned between these languages. However, for grade 7 it appears that the performance standard is lowest in Spanish, somewhat higher than that in German, and higher still in French. This situation could also be interpreted in terms of students needing to demonstrate a higher level of performance to achieve grade 7 in German than in Spanish, and higher than both of these in French.

An important caveat regarding the results for French in particular are the issues around the nature of the writing and speaking benchmarks, which makes the linking for French difficult to interpret for some grades. However, the ratings given post hoc by our panellists to these exemplars to some extent help with interpretation. Nevertheless, the results for French should be treated as tentative and needing corroboration, ideally using a different set of CEFR benchmarks.

# Standard linking of reading and listening comprehension assessments

## *Rater reliability analysis*

The analysis of the results of the standard linking task involved checking that the outcomes are based on reasonably reliable and consistent judgements. We report both agreement (i.e. showing agreement level in absolute terms) and consistency (i.e. showing agreement in terms of rank order of values) inter-class correlations (ICCs). An ICC score between 0.6 and 0.74 is usually considered good, and over 0.75 excellent (Cicchetti, 1994).

To check agreement and consistency levels of raters before discussion, and thus get a sense of how well they understood and were able to use the CEFR scales independently following the training that was provided, we calculated ICCs for initial ratings. We excluded the items which were used for familiarisation and initial discussion in the panels. Due to a technical problem, the initial ratings for French listening were unfortunately lost.

The table below shows the ICCs based on the initial ratings. It can be seen that even based on the initial ratings the ICCs were reasonably high. This suggests that our participants were able to use the CEFR scales independently with a high level of agreement and consistency.

Table 28 *ICCs based on initial ratings*

| Subject | ICC agreement | Confidence interval | ICC consistency | Confidence interval |
|---|---|---|---|---|
| Spanish reading | 0.81 | 0.70 < ICC < 0.89 | 0.83 | 0.74 < ICC < 0.91 |
| Spanish listening | 0.88 | 0.81 < ICC < 0.93 | 0.91 | 0.86 < ICC < 0.95 |
| German reading | 0.85 | 0.77 < ICC < 0.91 | 0.88 | 0.82 < ICC < 0.92 |
| German listening | 0.90 | 0.84 < ICC < 0.94 | 0.92 | 0.88 < ICC < 0.95 |
| French reading | 0.73 | 0.55 < ICC < 0.85 | 0.85 | 0.78 < ICC < 0.91 |

We also calculated the ICCs for the final ratings, which were given following the panel discussions and after the participants had the opportunity to amend their initial ratings – presented in the table below. The ratings for all tasks were included in this analysis. It can be seen that agreement and consistency levels of the ratings improved after discussion. The cut score and linking analysis presented in the following sections is based on these final ratings.

Table 29 *ICCs based on final ratings*

| Subject | ICC agreement | Confidence interval | ICC consistency | Confidence interval |
|---|---|---|---|---|
| Spanish reading | 0.89 | 0.84 < ICC < 0.93 | 0.90 | 0.86 < ICC < 0.94 |

| | | | | |
|---|---|---|---|---|
| Spanish listening | 0.97 | 0.96 < ICC < 0.98 | 0.97 | 0.96 < ICC < 0.98 |
| German reading | 0.98 | 0.97 < ICC < 0.98 | 0.96 | 0.97 < ICC < 0.98 |
| German listening | 0.95 | 0.93 < ICC < 0.97 | 0.96 | 0.94 < ICC < 0.97 |
| French reading | 0.95 | 0.93 < ICC < 0.97 | 0.96 | 0.94 < ICC < 0.97 |
| French listening | 0.91 | 0.87 < ICC < 0.94 | 0.92 | 0.88 < ICC < 0.95 |

## *Standard linking GCSE assessments to the CEFR*

### ***Reading comprehension***

Figures 32 to 34 show the distribution of ratings for each mark point in terms of sub-level ratings transformed into the sub-level numerical scale (left hand panel) and the distribution of ratings for each mark point after the mean sub-level ratings were transformed into the CEFR levels (right hand panel).

Recall that the sub-level ratings were transformed into the numerical scale shown in Table 9 in the Data analysis section. Once the sub-level ratings were transformed into the numerical scale, mean sub-level ratings for each mark point were calculated. These mean ratings were then transformed back into the CEFR level ratings, using the ranges shown in Table 10.

In Spanish, out of 50 marks, only one mark was deemed to be accessible to A1 level candidates. Twenty one  marks were deemed accessible to A2 candidates and 28 marks to B1 candidates. In German, out of 60 marks, one mark was deemed accessible to A1 level candidates. Thirty seven  marks were deemed accessible to A2 candidates and 22 to B1 candidates. In French, out of 60 marks, one mark was deemed accessible to A1 level candidates. Forty four  marks were deemed accessible to A2 candidates and 14 to B1 candidates.

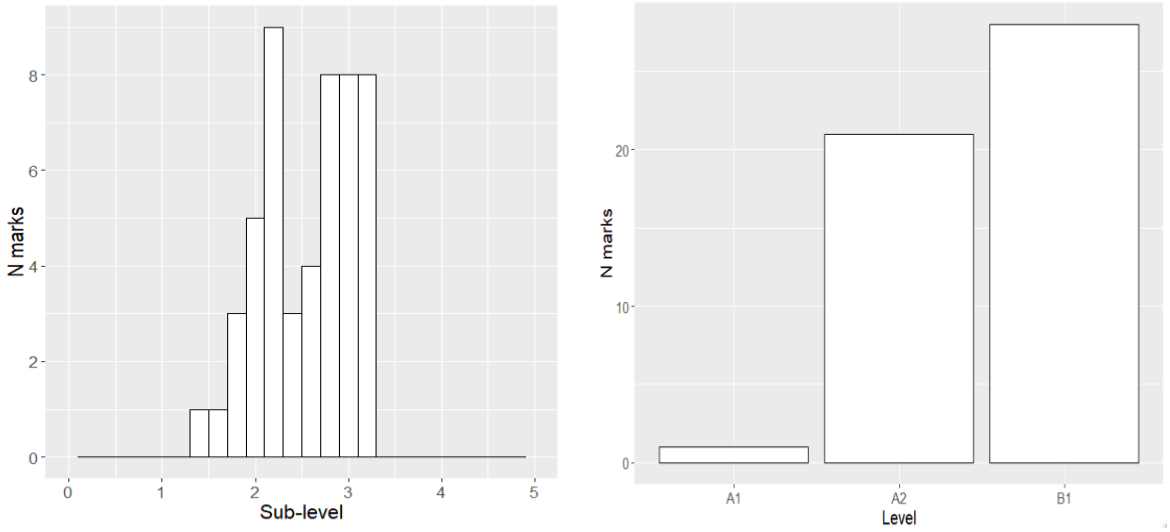


Figure 32 *Spanish reading comprehension - distribution of CEFR sub-levels and levels*

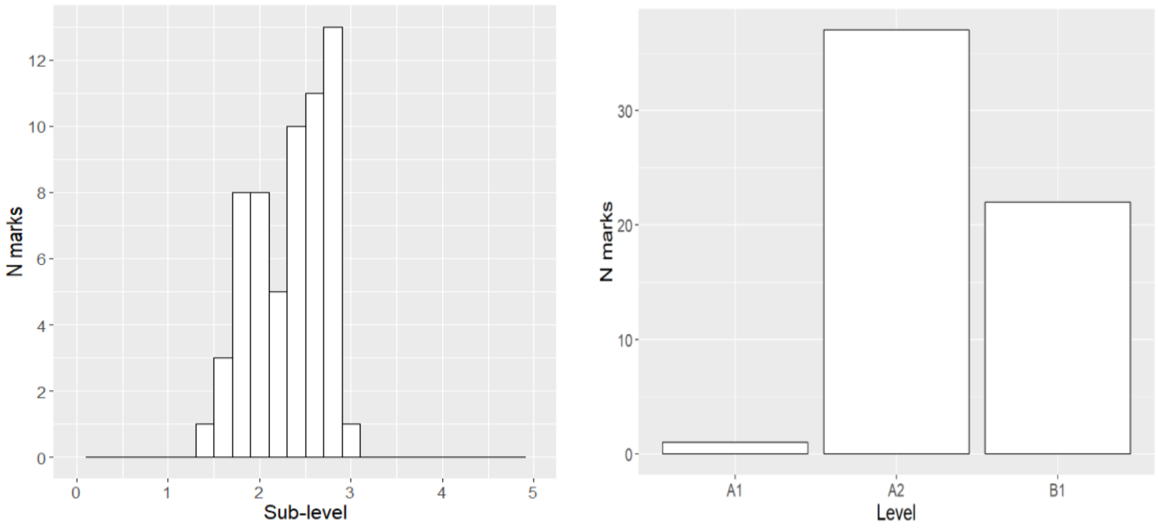


Figure 33 *German reading comprehension - distribution of CEFR sub-levels and levels*

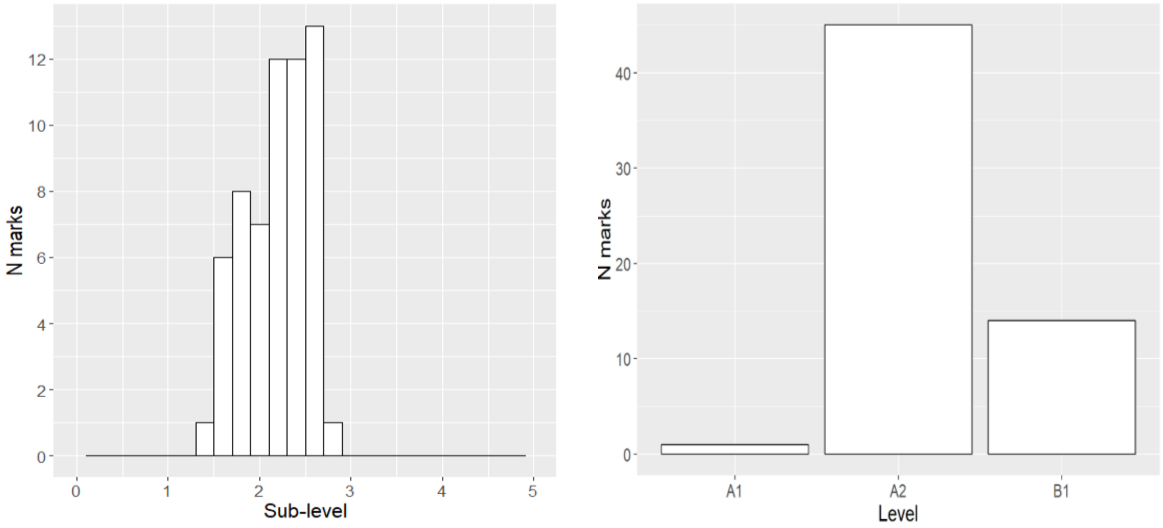


Figure 34 *French reading comprehension - distribution of CEFR sub-levels and levels*

The cut scores were set based on sums of weighted frequencies of the CEFR level ratings, as described in the Data analysis section. The table below shows the frequencies of mark points rated at different CEFR levels, and the resulting cut scores for each CEFR level and language.

Table 30 *CEFR level rating frequency and cut scores for reading*

| | Spanish | | German | | French | |
|---|---|---|---|---|---|---|
| CEFR level | N marks | Cut score | N marks | Cut score | N marks | Cut score |
| A1 | 1 | 1 | 1 | 1 | 1 | 1 |
| A2 | 21 | 12 | 37 | 20 | 44 | 23 |
| B1 | 28 | 32 | 22 | 42 | 14 | 44 |

According to these cut scores, in order to reach A2 level, a candidate would need to score 12 out of 50 marks for Spanish, 20 out of 60 marks for German and 23 out of

60 marks for French in the GCSE paper. In order to reach B1 level, a candidate would need to score 32 out of 50 marks in Spanish, 42 out of 60 in German and 44 out of 60 in French.

It was possible to set a cut score for A1 for each language as there was one mark point rated as accessible to A1 candidates in each test. However, based on content mapping, and the fact that this test is intended for Higher tier candidates, it would be more appropriate to conclude that any candidates scoring 11 or fewer marks would be deemed as A1 or below rather than those scoring just one mark.

The figures below provide an indication of how GCSE grades relate to the CEFR levels. In these figures, the CEFR cut scores are superimposed onto the GCSE mark distribution for the whole population of students that took the assessments in 2018 and the associated grade GCSE grade boundaries.

In Spanish, GCSE grade 4 might be described as low to mid A2 level, grade 7 as mid to high A2 level, and grade 9 as low to mid B1 level.

Therefore, the GCSE grades can be mapped approximately to the CEFR as follows:

Table 31 *GCSE to CEFR mapping for Spanish reading comprehension*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | Low-mid A2 | A2 |
| 7 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 |

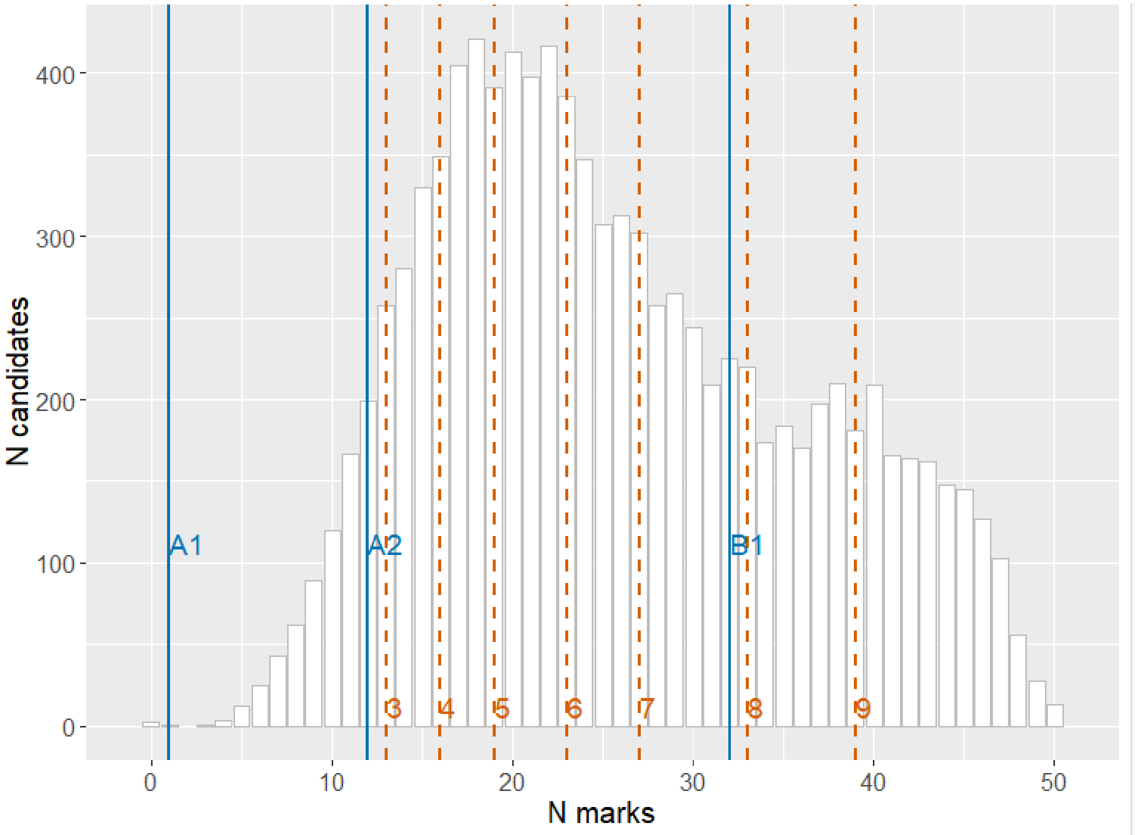


Figure 35 *Spanish reading comprehension – GCSE grade to CEFR mapping*

In German, GCSE grade 4 might be described as high A1 to low A2 level, grade 7 as mid to high A2 level, and grade 9 as low to mid B1 level.

Therefore, the GCSE grades can be mapped approximately to the CEFR as follows:

Table 32 *GCSE to CEFR mapping for German reading comprehension*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | High A1-low A2 | A1/A2 |
| 7 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 |

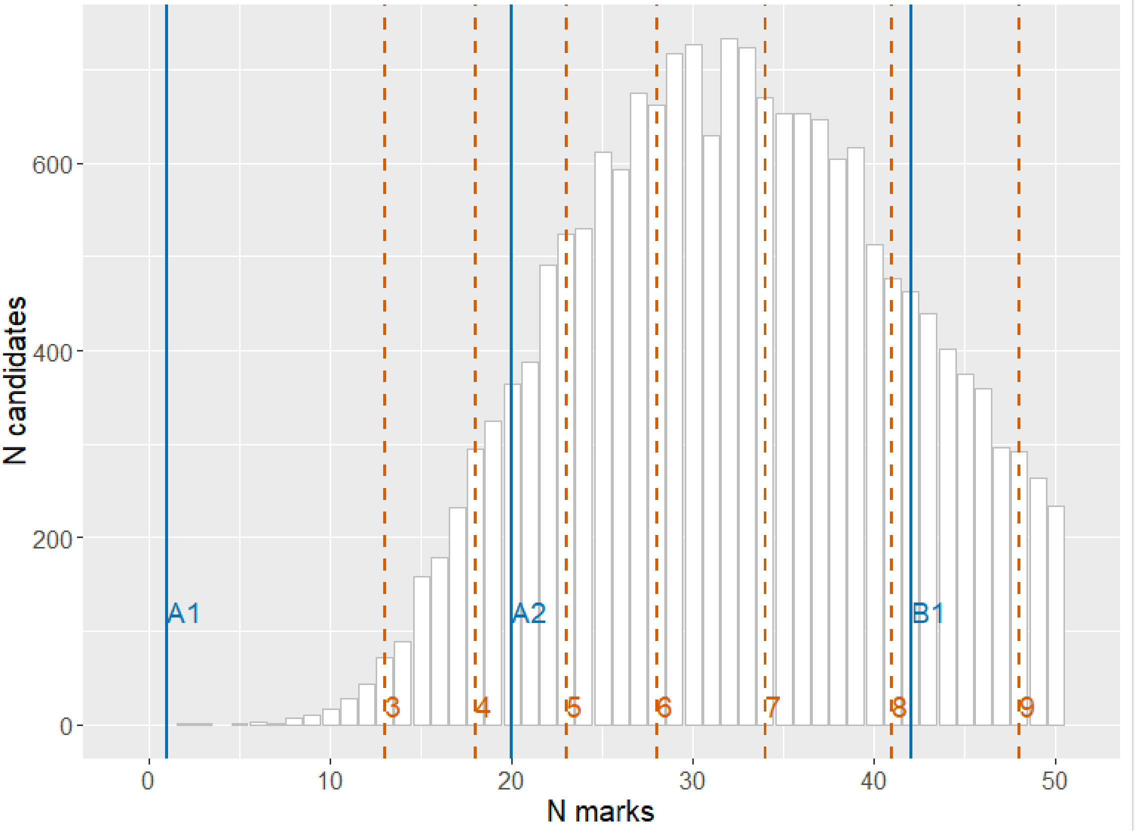


Figure 36 *German reading comprehension – GCSE grade to CEFR mapping*

In French, GCSE grade 4 might be described as high A1 to low A2 level, grade 7 as mid to high A2 level, and grade 9 as low to mid B1 level.

Therefore, the GCSE grades can be mapped approximately to the CEFR as follows:

Table 33 *GCSE to CEFR mapping for French reading comprehension*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | High A1-low A2 | A1/A2 |
| 7 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 |

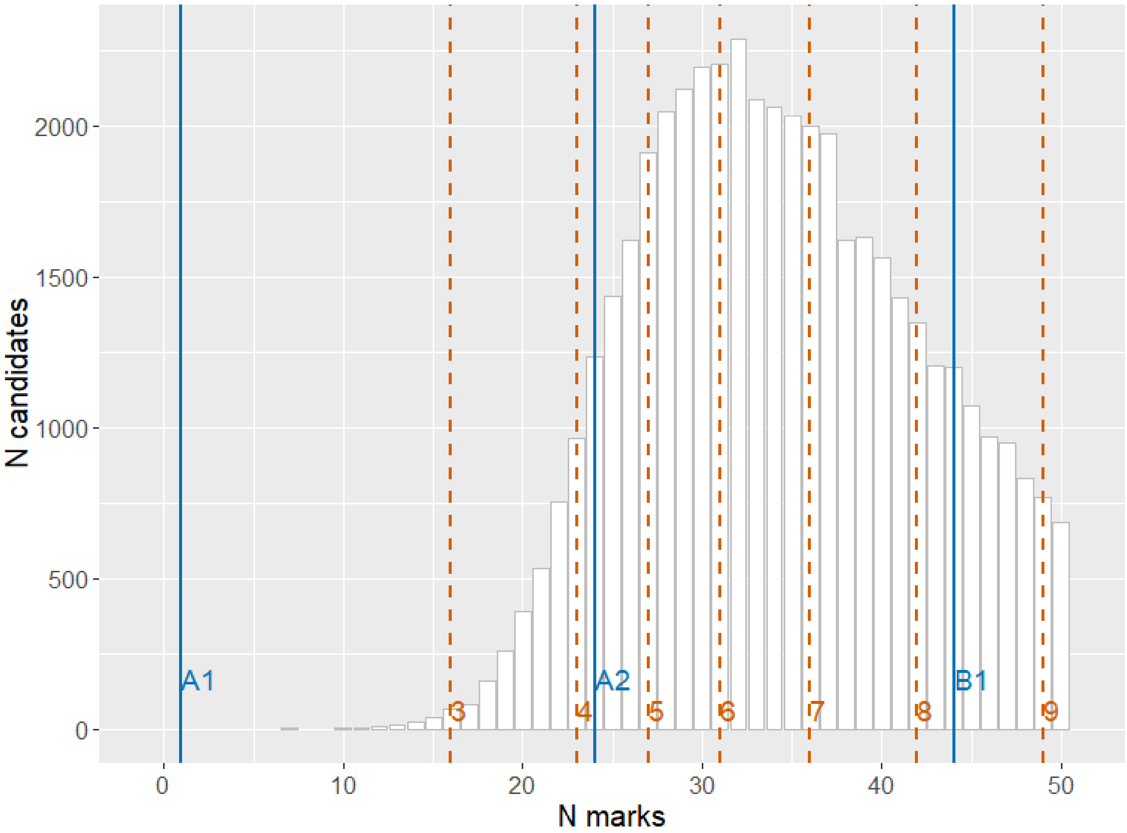


Figure 37 *French reading comprehension – GCSE grade to CEFR mapping*

## Listening comprehension

Figures 38 to 40 show the distribution of ratings for each task and mark point in terms of sub-level ratings transformed into the sub-level numerical scale and the distribution of ratings for each task and mark point after the mean sub-level ratings were transformed into the CEFR levels (cf. previous section).

The maximum mark for each listening paper is 50. In Spanish, 21 marks were deemed accessible to A2 candidates and 28 marks to B1 candidates. One mark was deemed accessible to B2 candidates. In German, one mark was deemed accessible to A1 level candidates. 38 marks were deemed accessible to A2 candidates and 11 to B1 candidates. In French, out of 60 marks, one mark point was deemed accessible to A1 level candidates. 44 marks were deemed accessible to A2 candidates and 14 to B1 candidates.

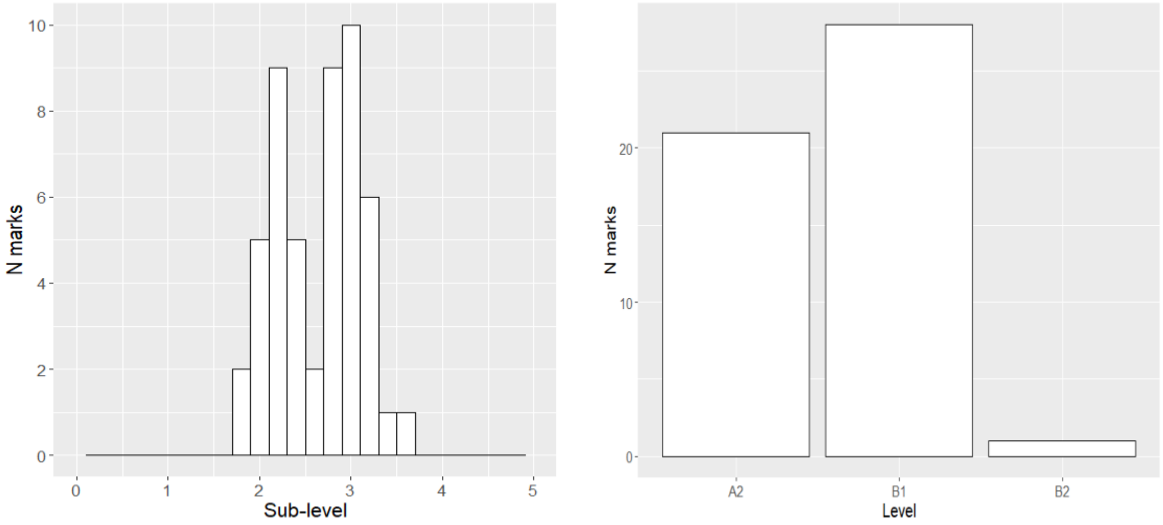


Figure 38 *Spanish listening comprehension - distribution of CEFR sub-levels and levels*

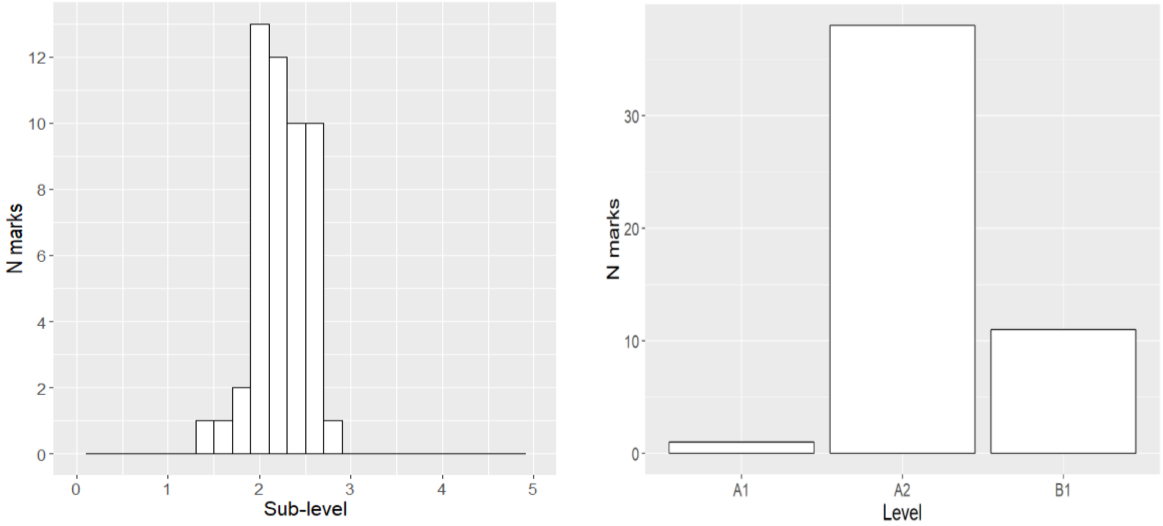


Figure 39 *German listening comprehension - distribution of CEFR sub-levels and levels*

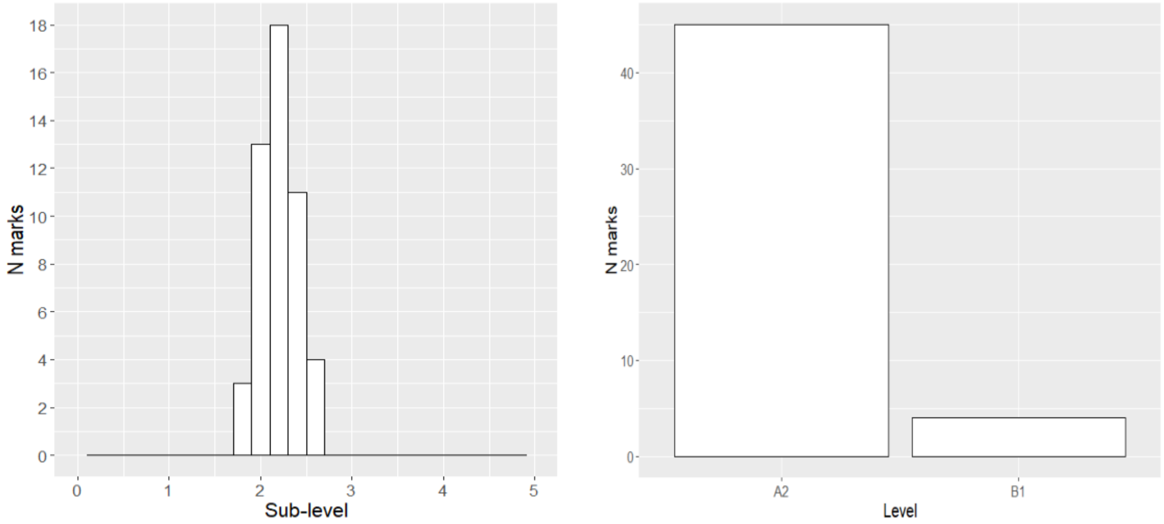


Figure 40 *French listening comprehension - distribution of CEFR sub-levels and levels*

The table below shows the frequencies of mark points rated at different CEFR levels, and the resulting cut scores for each level.

Table 34 *CEFR level rating frequency and cut scores for listening*

| | Spanish | | German | | French | |
|---|---|---|---|---|---|---|
| CEFR level | N marks | Cut score | N marks | Cut score | N marks | Cut score |
| A1 | | | 1 | 1 | | |
| A2 | 21 | 11 | 38 | 20 | 45 | 23 |
| B1 | 28 | 31 | 11 | 37 | 4 | 38 |
| B2 | 1 | 43 | | | | |

According to these cut scores, in order to reach A2 level, a candidate would need to score 11 marks in Spanish, 20 marks in German and 23 marks in French out of 50 marks in the GCSE paper. In order to reach B1 level, a candidate would need to score 31 marks in Spanish, 37 marks in German and 38 marks in French.

It was possible to set a cut score for A1 for German and for B2 in Spanish. However, similarly to reading comprehension, based on content mapping, and the fact that this test targets Higher tier candidates, it would be more appropriate to conclude that any candidates scoring 19 or fewer marks in German would be deemed as A1 or below rather than those scoring just one mark. Similarly, candidates scoring 32 and higher might be more appropriately considered to be of B1 level.

In Spanish, GCSE grade 4 might be best described as low to mid A2 level, grade 7 as mid to high A2 level, and grade 9 as mid to high B1 level.

Therefore, the GCSE grades can be mapped approximately to the CEFR as follows:

Table 35 *GCSE to CEFR mapping for Spanish listening comprehension*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | Low-mid A2 | A2 |
| 7 | Mid-high A2 | A2 |
| 9 | Mid-high B1 | B1 |

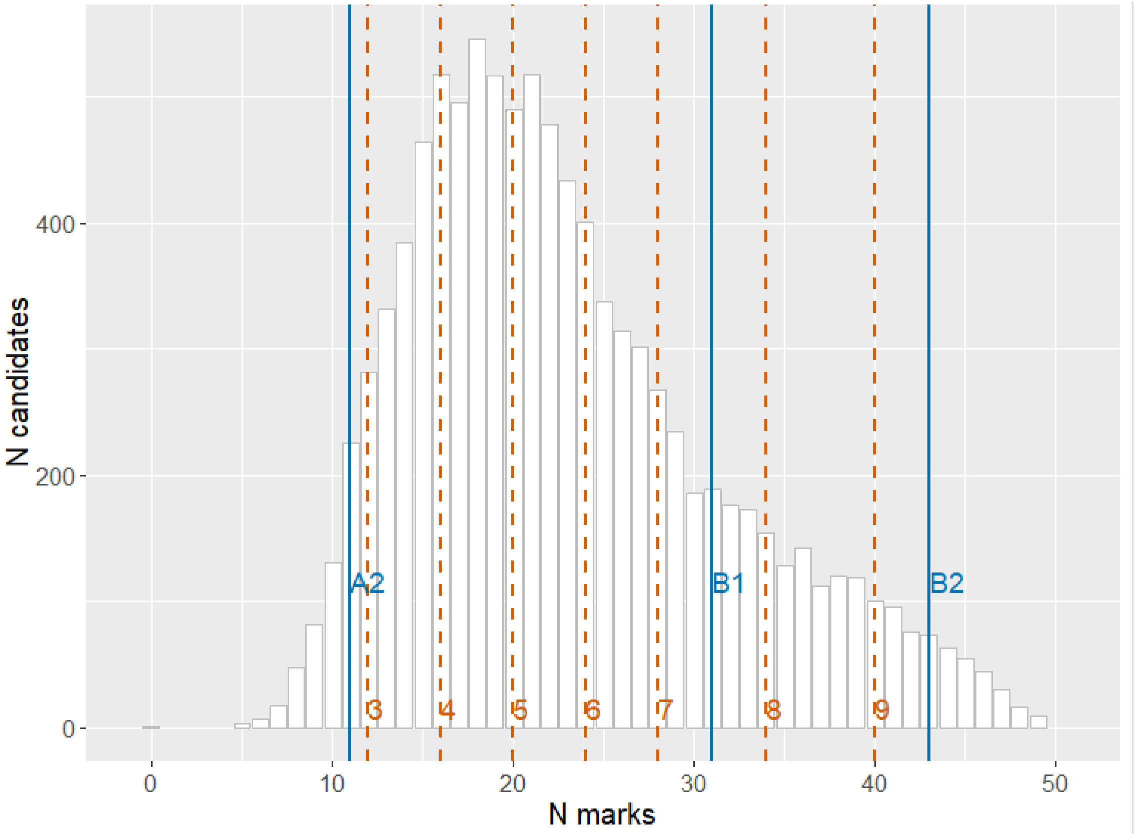


Figure 41 *Spanish listening comprehension – GCSE grade to CEFR mapping*

In German, GCSE grade 4 might be best described as high A1 to low A2 level, grade 7 as mid to high A2 level, and grade 9 as low to mid B1 level.

Therefore, the GCSE grades can be mapped approximately to the CEFR as follows:

Table 36 *GCSE to CEFR mapping for German listening comprehension*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | High A1-low A2 | A1/A2 |
| 7 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 |

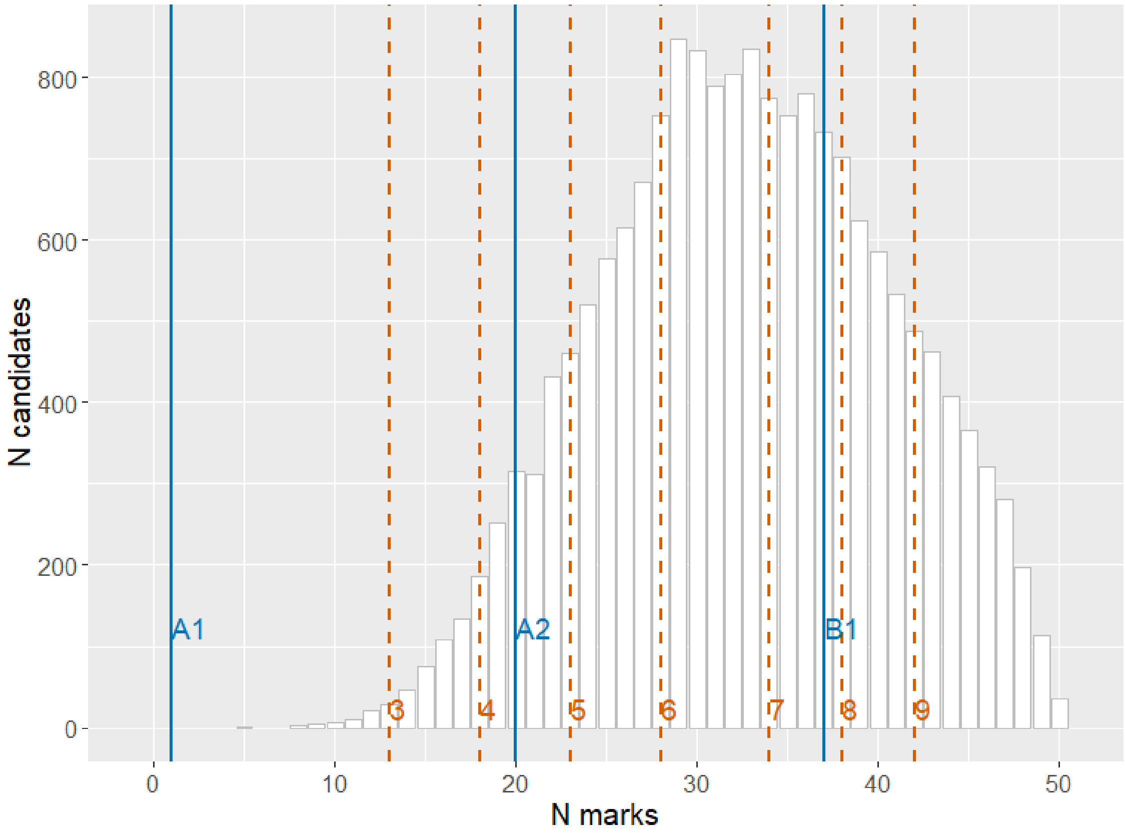


Figure 42 *German listening comprehension – GCSE grade to CEFR mapping*

In French, GCSE grade 4 might be best described as A1 level, possibly low-mid A1. Grade 7 might be described as high A1 to low A2 level, and grade 9 as high A2 to low B1 level.

Therefore, the GCSE grades can be mapped approximately to the CEFR as follows:

Table 37 *GCSE to CEFR mapping for French listening comprehension*

| GCSE grade | CEFR sub level | CEFR level |
|---|---|---|
| 4 | Low-mid A1 | A1 |
| 7 | High A1-low A2 | A1/A2 |
| 9 | High A2-lowB1 | A2/B1 |

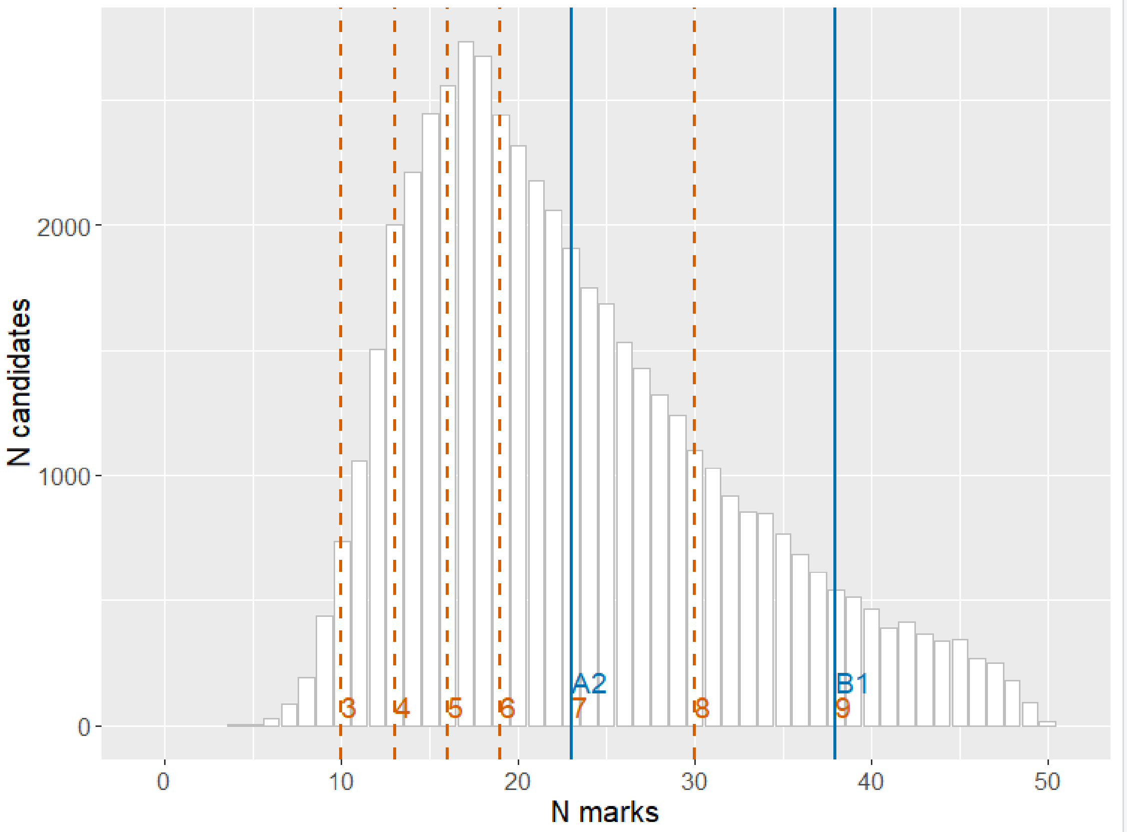


Figure 43 *French listening comprehension – GCSE grade to CEFR mapping*

## Summary and interim discussion

The tables below summarise the GCSE to CEFR linking for receptive skills.

Table 38 *GCSE to CEFR mapping for Spanish receptive skills*

| | Reading | | Listening | |
|---|---|---|---|---|
| GCSE grade | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level |
| 4 | Low-mid A2 | A2 | Low-mid A2 | A2 |
| 7 | Mid-high A2 | A2 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 | Mid-high B1 | B1 |

Table 39 *GCSE to CEFR mapping for German receptive skills*

| | Reading | | Listening | |
|---|---|---|---|---|
| GCSE grade | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level |
| 4 | High A1-low A2 | A1/A2 | High A1-low A2 | A1/A2 |
| 7 | Mid-high A2 | A2 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 | Low-mid B1 | B1 |

Table 40 *GCSE to CEFR mapping for French receptive skills*

| | Reading | | Listening | |
|---|---|---|---|---|
| GCSE grade | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level |
| 4 | High A1-low A2 | A1/A2 | Low-mid A1 | A1 |
| 7 | Mid-high A2 | A2 | High A1-low A2 | A1/A2 |
| 9 | Low-mid B1 | B1 | High A2-lowB1 | A2/B1 |

Within each language, the reading and listening papers are quite consistent in terms of the mapping of GCSE grades to CEFR levels, and in terms of the percentage of total marks required to achieve each CEFR level. This is shown in the table below.

Table 41 *Percentage of total marks required for each CEFR level*

| | Reading | | Listening | |
|---|---|---|---|---|
| Subject | A2 | B1 | A2 | B1 |
| Spanish | 24 | 64 | 22 | 62 |
| German | 33 | 70 | 40 | 74 |
| French | 38 | 73 | 38 | 63 |

However, French listening paper differs significantly from French reading paper in that GCSE grades map to lower CEFR levels for listening. The table below shows that, in terms of proportion of total marks required for different GCSE grades, French listening paper has a different profile compared to French reading paper, as well as compared to the listening papers in other languages.

Table 42 *Percentage of total marks required for each GCSE grade*

| | Reading | | | Listening | | |
|---|---|---|---|---|---|---|
| Subject | 4 | 7 | 9 | 4 | 7 | 9 |
| Spanish | 32 | 54 | 78 | 32 | 56 | 80 |
| German | 30 | 57 | 80 | 36 | 68 | 84 |
| French | 38 | 60 | 82 | 26 | 46 | 76 |

This suggests that there may have been something specific to either the French listening paper[24] or the French cohort of students, or both, that has affected the performance of students on this test, and prevented them from scoring at the same level as their peers did on the other 2 listening papers. The standard-maintaining approach currently implemented for GCSEs ensures that candidates in 2018 are, on average, as likely as candidates in 2017 showing similar prior attainment to achieve a particular grade. Therefore, the grade boundaries for the French listening paper may be lower compared to German and Spanish listening papers to accommodate apparent higher difficulty of the French paper.

In terms of the CEFR mapping, however, this situation means that the students achieving grade 4 on French listening paper effectively demonstrated a substantively lower level of performance than on the other papers considered here. To the extent that this may have been due to issues with unintended sources of difficulty in the test itself, this needs to be addressed in test development in order to ensure that the papers are set at appropriate levels of demand, while including fewer unintended sources of difficulty (cf. Pollitt et al., 2008).

Straightforward direct comparisons across the 3 specifications cannot be made as the linking is based on the judgements from different panels. However, on the assumption that the content and examination standards, as well as test specifications, are supposed to be reasonably aligned in these 3 languages, it may

[24] It is difficult to pinpoint exactly why this may have been the case, but research carried out by Stratton and Zanini (2019) suggests that a higher proportion of constructed response (CR) items in the 2018 French listening assessments for higher tier compared to other languages may have increased their difficulty. Their analysis was carried out across exam boards rather than just for this particular specification. However, for the current specification there were only 9 CR items in 2017 compared to 21 in 2018. In addition, compared to 21 CR items in French listening, there were only 12 and 15 in the 2018 German and Spanish listening papers respectively (8 and 7 in 2017 respectively). This could go some way towards explaining the apparent higher difficulty of the French listening paper, especially when considered in conjunction with the comments made in the French panel that the mark scheme for the CR items was often too specific or restrictive, which could have more of an impact on overall paper difficulty when there is a higher proportion of CR items. If such mark scheme features were unintended or not construct-relevant, these issues could have contributed to test difficulty without affecting the intended demand of the test. This could account for the fact that the CEFR level ratings, which are primarily related to intended task demand, being are similarly distributed for all receptive skills tests, even where, like in French listening, the actual test performance of candidates may have been affected by unintended sources of difficulty, which would not have been recognised in the CEFR ratings.

Stratton and Zanini also found a significant effect on test difficulty of higher speech speed for French listening higher tier exams across all specifications (again the analysis was not done separately for each exam board). This may have additionally contributed to the difficulty of the AQA French listening paper compared to the other listening papers considered here. It should be noted that speech speed was not mentioned as problematic by the French panel and it was deemed appropriate for the overall level of demand. However, the speech speed in the other 2 languages was characterised as non-standard by our consultant.

be considered plausible and informative to consider the results of the linking across the 3 languages.

Except for the French listening paper, the other papers are quite consistent in terms of performance levels required for grades 7 (mid-high A2) and 9 (low B1) and thus in terms of the CEFR levels. Performance standard required for Grade 4 in Spanish appears to be higher in both reading and listening papers than in the other 2 languages. This is consistent with a higher proportion of tasks rated as B1 in Spanish compared to the other 2 tests.

There are some caveats to the interpretation that the standard for Spanish at grade 4 is higher than in the other 2 languages. Firstly, it should be noted that the German and French performances are based on the tests from the same exam board, whereas the Spanish test came from a different board. While the accreditation of these specification required them to be comparable, accreditation itself could not entirely guard against inter-board differences. Thus, the difference with respect to grade 4 may be related to the demand of this specific test and associated assessment criteria rather than overall Spanish test demand and Spanish performance standard across exam boards. On the other hand, on the assumption that this Spanish specification is well aligned with the other Spanish specifications, another interpretation of this finding could be related to the fact that Spanish candidates are expected to be more able than the candidates in other languages, and hence able to deal with higher task demand even at grade 4. This could justify a higher performance standard and higher test demand across different Spanish specifications. In this scenario, however, the meaning of grade 4 in Spanish would be different than the meaning of grade 4 in German, for example, as they would represent different performance standards.

# Qualitative results

In this section, we summarise the main themes that emerged from the discussions carried out with the 3 standard-linking panels. The analysis highlighted a variety of themes which can broadly be placed into three categories: “Subject content”, “Nature of assessment” and “The purpose of MFL GCSE”. Additionally, the concepts of spontaneity[25] (e.g. Mitchener, 2016), interaction and communicative language skills were emphasised throughout.

## *Subject content*

The content of MFL GCSEs was one of the more prominent themes discussed. Conversations mainly focussed around the range of topics and grammatical structures the curriculum currently covers and the way teaching and assessment of grammar interacts with teaching and assessment of communication.

Panellists discussed how the range and variety of topics prevented in-depth learning, and were in some cases wider than what might be expected at comparable levels in the CEFR. This may have the added consequence of inhibiting the development of independent inquiry skills, necessary for study at higher levels.

> *“I’m used to calling it greater depth, but it’s … in-depth learning. That’s a feature of the national curriculum right, from key stage 2 and at key stage 3, and it’s not a feature of GCSE. GCSE is about racing through content to score marks, and they’re totally at odds with each other.” [French panel participant]*

> *“There are too many topics, and teachers feel extremely under pressure … teachers feel they have to go through all of these topics. You’ve got no time at all, even when you’ve got few hours per week, to really get into the culture” [French panel participant]*

> *“They did have to talk about a wider range of subjects than in the CEFR.” [German panel participant]*

> *“At GCSE level if you say to them sorry we don’t have time. Yeah, very interesting question but you know what, we don’t have time to answer this, and we’re focusing on this. So they get to the end of their A-level, and you say to them think for yourself, and they go no, tell me what I’ve got to do. Because the whole of my curriculum you’ve told me what I do to pass this exam.” [French panel participant]*

Additionally, some panellists expressed the view that some of the topics covered are outdated or irrelevant, possibly presenting a distraction to students and deterring engagement in MFL subjects.

> *“And … don’t talk about teacher training, or the fact that your mum’s injecting herself and taking drugs, or some of the horrendous topics that are quite frankly, if I was 14 would put me off.” [French panel participant]*

Panels also discussed the level of complexity at which grammar should be taught, suggesting that a lower level of complexity might be better to allow more in-depth

[25] Spontaneity refers to the ability of students to communicate effectively and naturally with native language users. Both spontaneity and communicative language skills are terms used in the CEFR to categorise language ability.

understanding and productive use. The range of structures introduced was also noted as being too broad for the level that might be expected at GCSE.

> *"I would much rather they had less grammatical complexity at GCSE, … I would much rather than the whole linguistic level came down, but they really understood it, and knew it and could use it." [French panel participant]*
>
> *"I think you've got to take it a sentence at a time and every sentence has complications in it. They have subordinate clauses, they have past modals, future tenses" [German panel participant]*
>
> *"I don't think that the kind of task is testing so much on the comprehension of the text. It's also structure and language. So you need to understand if you need to use a subjunctive, an indicative test, so it's more about language structure I think than the actual topic." [Spanish panel participant]*
>
> *"Sometimes it was hard because I was listening thinking well they definitely sound much more fluent, but what they're saying is simpler grammatically. But then to communicate, they're communicating so much better. And I think that's where you need to do something" [French panel participant]*

Despite being critical about the range and complexity of the grammatical structures in the curriculum, panellists across all 3 subjects seemed to think that explicit teaching of grammar in the classroom was necessary for achieving productive knowledge of language, as well as for students who wish to study at a higher level. In particular, panellists from the German panel discussed at length the benefits of teaching grammar in order for students to be able to use language independently in a flexible manner.

> *"If you move away from grammar how do you teach language; you have to learn it parrot fashion and if you have the grammar then you're flexible." [German panel participant]*
>
> *"With teaching a little bit more grammar you also learn to make your own sentences so you become more independent on your own." [German panel participant]*
>
> *"When we get them in the first year of university they don't know the first thing about grammar, they don't have the words." [French panel participant]*

One panel member from the German panel also described the increased focus on grammar as a major improvement of the new GCSE.

> *"What I like about the new GCSE is that we have more permission to work with grammar … so we've had the permission to do more grammar and you can actually just prepare the children just to become more independent" [German panel participant]*

This view was qualified by views from other panellists, who felt that there needs to be a balance between teaching and assessing communicative ability vs. precision in grammar use. They felt that some aspects of assessment (e.g. speaking exam) should reward effective communication rather than penalise absence or imprecise use of certain complex grammatical structures at this level.

> *"When they come [to the exam, and] they're so communicative … and then the grammar is like hmm there's no subordination. But at the end of the day*

> *it's an oral exam. So if they communicate, I think that should be [the point]." [Spanish panel participant]*

> *"They were thinking mistakenly of raising grammatical standards and raising linguistic standards… not thinking about raising communication standards…all they've done is they've made … better grammatical linguists, more analytical linguists in that respect. But we're making worse practical linguists." [French panel participant]*

This, and other issues related to nature of assessment are discussed in more detail in the next section.

## *Nature of assessment*

Discussion in the panels also covered a variety of themes relating to the nature of GCSE assessment. This category can broadly be split into3 themes, which relate to the appropriateness/effectiveness of exams, the "trickery"/ deception of some question formats, and the instances of negative/harsh marking.

The first of these themes relates to effective assessment of skills which the panellists felt should be achieved following completion of an MFL GCSE. Certain exams were felt to provide better opportunities to display desirable linguistic skills. For example, some panellists agreed the current writing exam presents a good opportunity to display spontaneous language skills due to the inability to pre-learn responses.

> *"I think with the writing actually that's where you get the most evidence of spontaneity, because actually it is difficult to pre-learn for the writing … you do have no idea, you have no input about what topics you're getting for the writing; you have very limited stimulus that might just say here's 2 bullet points or whatever. So I think actually in terms of spontaneity the writing is the best place they can show that really." [French panel participant]*

In contrast to this, the speaking exam was highlighted as an area where less spontaneity, or even 'faked spontaneity' may be demonstrated, and where there was little assessment of interaction:

> *"It's more or less a monologue with some teacher interference." [German panel participant]*

> *"They don't sound natural or spontaneous … it still feels that they've gone through loads of questions, they've practised all the answers, so they kind of know what they're going to say to every question [German panel participant]*

> *"They do need speaking skills, but they're not going to have the chance to learn things by heart. They need to be able to have a conversation, and the way the current speaking exam is organised they're able to learn a piece by heart." [French panel participant]*

> *"[The GCSE speaking exam] that's not really a conversation, I think speaking, that's rehearsing." [German panel participant]*

> *"In GCSE there's not real interaction. So one says a sentence, but then the question hasn't, is not linked to what the student has answered. So there's really no interaction." [Spanish panel participant]*

> *"I wasn't convinced with this interaction though. They were taking turns, but there was no probing or pushing … it seemed that most of what they said was pretty pre-prepared and predictable. There was no moment where someone said tell me more, or give me an example." [Spanish panel participant]*

The second theme refers to what some panellists described as "trickery" employed in exams. This is where the layout or other properties of questions within exam papers may make it more difficult to arrive at a correct response and therefore such features potentially act as construct-irrelevant source of difficulty. While in some cases it is possible that particular odd or novel question formats are introduced in tests deliberately to test for specific aspects of knowledge, detailed comprehension, etc., it is important that these do not simultaneously represent construct-irrelevant variance. Some of our panellists suggested that they saw evidence of construct-irrelevant variance in these tests and were of the view that this sometimes also got in the way of students demonstrating their actual linguistic skills.

> *"It seems that you have to prepare students for 2 things. One is Spanish at that level, and another thing is OK, this is an obstacle course, and this is what they're going to throw at you. And you're going to have to do this, that and the other to get to the end." [Spanish panel participant]*

> *"If something is labelled A, B, C, then you expect it to be in that order." [Spanish panel participant]*

> *"We teach the students in any institute or in GCSE level, we always teach them that it's in order, questions are always in order. So the fact that [the next answer in the text] is before [the last answer] they're not going to even look before." [Spanish panel participant]*

> *"I feel the problem is with this task you can't find, given the text and the answers you can't actually find out what the students understand. You actually do test their world knowledge… If they know the fairy tale then it's actually quite easy, 'oh yeah, I remember that', but if you don't know it then you actually have to read the text." [German panel participant]*

Additionally, some specific examples were discussed where voice actors in the listening exam may have used a misleading tone of voice:

> *"…sometimes [the speakers] try to trick you that they sound really sad, but are really happy about something" [German panel participant]*

> *"[The speaker] was just talking as if she was really angry and in a very loud voice. And that wasn't really the purpose of what she was saying, but then immediately that's the reaction of the student" [Spanish panel participant]*

The third theme refers to what panellists described as instances of harsh and negative mark schemes. Some panellists felt this may prohibit interactive and spontaneous communication skills, as marking was very prescribed, not allowing room for contextually correct answers to be accepted.

> *Participant 1: "People* [test developers] *have sat there and thought how can we catch them."*
> *Participant 2: "And that's very much reflected in the mark scheme because*

*they're after very tight answers. And then that impacts obviously on the way you distribute marks towards the different grades."*
*Participant 3: "And again it could hamper someone more able because they won't have necessarily the same standardised answers." [French panel participants]*

*"I always feel as a native speaker you are always a bit harsher with everything, but there I just felt I would have given them that point." [German panel participant]*

*Participant 1: "Bring the level of trickery down… and reward what they can do."*
*Participant 2: "Yes… positive marking." [French panel participants]*

*"Make it about what they can do, and reward what they can do, rather than penalising." [French panel participant]*

*"I talk about linguistic skills that I think you should look out for. The ability to rephrase, the ability to manipulate language, which I'm not sure still are rewarded sufficiently."[German panel participant]*

*"… change your marking criteria and you'll see a big difference ... Because actually that in itself will be enough to motivate them, because it won't be about oh I'm going to do this little trick, and this little trick. And more about OK how do I understand, how do I? And they'll start enjoying it I think." [French panel participant]*

## *The purpose of MFL GCSE*

In raising various issues and thinking about how current GCSE MFL assessments might be improved, panellists seemed to express uncertainty around the intended purpose of MFL GCSEs. It was felt that a clearer understanding of what students' language skills should look like on completion of MFL GCSE's was needed in order to be able to address any issues and make the qualifications more effective.

*"There's lots of [emphasis of] academic exercise over communication. I mean the thing is probably from step one is … what do we want a GCSE to be? And that's the starting point." [French panel participant]*

*"…language is expected to help them with their communication skills in the wider world, and that's precisely what the GCSE does not do." [French panel participant]*

*"What you are going to look like, as a learner, as a person … at the end, so just to envisage that, because our learners come out with such low self-confidence." [German panel participant]*

*"I think if you could change one thing, the one thing I would change is start really bottom up. What do we want them to be? We want them to be able to communicate. Right, where do we take the qualification from there?" [French panel participant]*

*"And it would be so much better if everything at GCSE was about giving people a skill" [French panel participant]*

Furthermore, panellists described a lack of continuity from GCSE to higher levels of study resulting from unclear intended outcomes.

> *"So there is a massive jump from GCSE to A-levels, and so the qualification is not preparing people."[French panel participant]*
>
> *"To have some way of having a better continuum through from key stage 2 into 3 as well, and right the way through because it doesn't follow through at the moment." [French panel participant]*
>
> *"The whole thing needs to be rethought, because then what are they going to do at A-level?" [Spanish panel participant]*

Panellists expressed further concerns about the lack of confidence even the highest achieving students have in using language in real-world situations. This consequence was attributed to a lack of practice of interactive and communicative language.

> *"[The students] have no confidence… I suspect largely because of the learning by rote … and also they have no exposure to German speakers here ever, so then to suddenly go to Germany and have to speak German, they're scared to do it." [German panel participant]*
>
> *"They learn very much exam techniques, and knowing the keywords that often are used to trip up. … You can really know those and do better in this than you perhaps would if you went to France and actually tried to speak." [French panel participant]*
>
> *"I've hardly seen any English students being able in Germany to go out, speak with a friend, order a pizza." [German panel participant]*
>
> *"What you are going to look like, as a learner, as a person, look like at the end, so just to envisage that, because our learners come out with such low self-confidence." [German panel participant]*

## *Summary and interim discussion*

The discussions with panellists revealed clear doubt around the effectiveness of current MFL GCSEs to instil the ability and confidence in learners to communicate with other language users in real-world context. Moreover, panellists highlighted how students who do go on to study at a higher level appear to be underprepared, necessitating time spent on getting them to the expected standard to study for A level or a university degree.

Possible reasons for this relate to the necessity to focus teaching on exam-specific material, covering a broad range of topics relatively superficially. This may restrict teachers in teaching grammar in a way which is helpful for spontaneous, creative and communicative language use. It was noted that students were required to know a great range of grammatical structures, but their productive use of these was not effectively instilled or assessed, particularly in speaking assessments.

Other prominent concerns were raised regarding the restrictive mark schemes which seem to penalise students for deviating from overly precise responses. This was particularly noted in relation to speaking assessments, but also in translation and other tasks, for instance, in French and German listening comprehension. This, coupled with some exam question formats, apparently invites responses of a rehearsed nature, rather than responses demonstrating interactive communication and spontaneous language skills or genuine language comprehension.

The concept of spontaneity – the ability of language users to communicate with and respond to native speakers effectively in a natural and relevant manner – is a key term used throughout the CEFR, and was repeatedly brought up during the panels. Panellists described current MFL GCSE teaching and assessment as under-emphasising (or even undermining) spontaneous language use, which would improve students' ability to interact with other language users and communicate effectively and confidently. Overall, this may have the consequence that students perceive MFL GCSEs as unbeneficial and de-motivating, as they do not gain a skill which is immediately useful in a real-world context.

# Discussion

In this section, we bring together the linking outcomes for all components of each specification. We discuss the implications of the linking for the grading standards as well as wider implications our findings have for further attempts at linking the CEFR and GCSE MFLs, and for assessment practice.

Tables 43-45 below present the GCSE to CEFR linking at component level for each language. As already observed, the linking of GCSE grades to the CEFR levels across components within Spanish and German is very consistent, with productive skills being at a lower CEFR level than the receptive skills. French mapping is less consistent, but this may be partly due to the issues with the CEFR exemplars for productive skills, and apparent issues with the listening comprehension paper. Therefore, we would suggest that the linking for French is more tentative than for the other 2 languages.

The patterns are broadly consistent across the 3 languages, with the notable exception of grade 7 for productive skills (lowest standard in Spanish), and grade 4 for receptive skills (highest standard in Spanish). While a degree of consistency is perhaps not surprising, and probably suggests that specification and assessment demands are generally comparable across different languages, there is no particular reason why we should expect the performance standards to be perfectly aligned across languages. There might be valid reasons why performance standards in one language may be deliberately higher than in another language. There are inevitably some intrinsic differences in difficulty between different languages for first language English speakers. For example, it is generally considered that Spanish and French are among the easier languages for English native speakers to acquire. Therefore, it might be reasonable to expect a higher level of performance (at any grade) in these languages compared to, perhaps, German.[26] As another example, if perhaps the students are considered to start from a higher level of ability in a language at the start of a GCSE course due to prior learning, this could explain higher performance standards at GCSE for this language. Whether or not this is then explicitly implemented in the nature of assessment is another matter, as there may be other reasons why requiring higher standards in one language compared to another may not be desirable. These issues remind us that considering standards between even quite related subjects involves considerable nuance and interpretation.

[26] https://www.atlasandboots.com/foreign-service-institute-language-difficulty/

Table 43 *GCSE to CEFR mapping for Spanish*

| | Writing | | Speaking | | Reading | | Listening | |
|---|---|---|---|---|---|---|---|---|
| GCSE grade | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level |
| 4 | Mid-high A1 | A1 | Low-mid A1 | A1 | Low-mid A2 | A2 | Low-mid A2 | A2 |
| 7 | Low-mid A2 | A2 | Low-mid A2 | A2 | Mid-high A2 | A2 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 | Low-mid B1 | B1 | Low-mid B1 | B1 | Low-mid B1 | B1 |

Table 44 *GCSE to CEFR mapping for German*

| | Writing | | Speaking | | Reading | | Listening | |
|---|---|---|---|---|---|---|---|---|
| GCSE grade | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level |
| 4 | Low-mid A1 | A1 | Mid A1 | A1 | High A1-low A2 | A1/A2 | High A1-low A2 | A1/A2 |
| 7 | Mid-high A2 | A2 | High A2 | A2 | Mid-high A2 | A2 | Mid-high A2 | A2 |
| 9 | Low-mid B1 | B1 | Low B1 | B1 | Low-mid B1 | B1 | Low-mid B1 | B1 |

Table 45 *GCSE to CEFR mapping for French*

| | Writing | | Speaking | | Reading | | Listening | |
|---|---|---|---|---|---|---|---|---|
| GCSE grade | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level | CEFR sub-level | CEFR level |
| 4 | High A1-Low A2 | A1/2 | Low-mid A1 | A1 | High A1-low A2 | A1/A2 | Low-mid A1 | A1 |
| 7 | Low-mid B1 | B1 | High A2-low B1 | A2/B1 | Mid-high A2 | A2 | High A1-low A2 | A1/A2 |
| 9 | Low-mid B1 | B1 | Mid-high B1 | B1 | Low-mid B1 | B1 | High A2-lowB1 | A2/B1 |

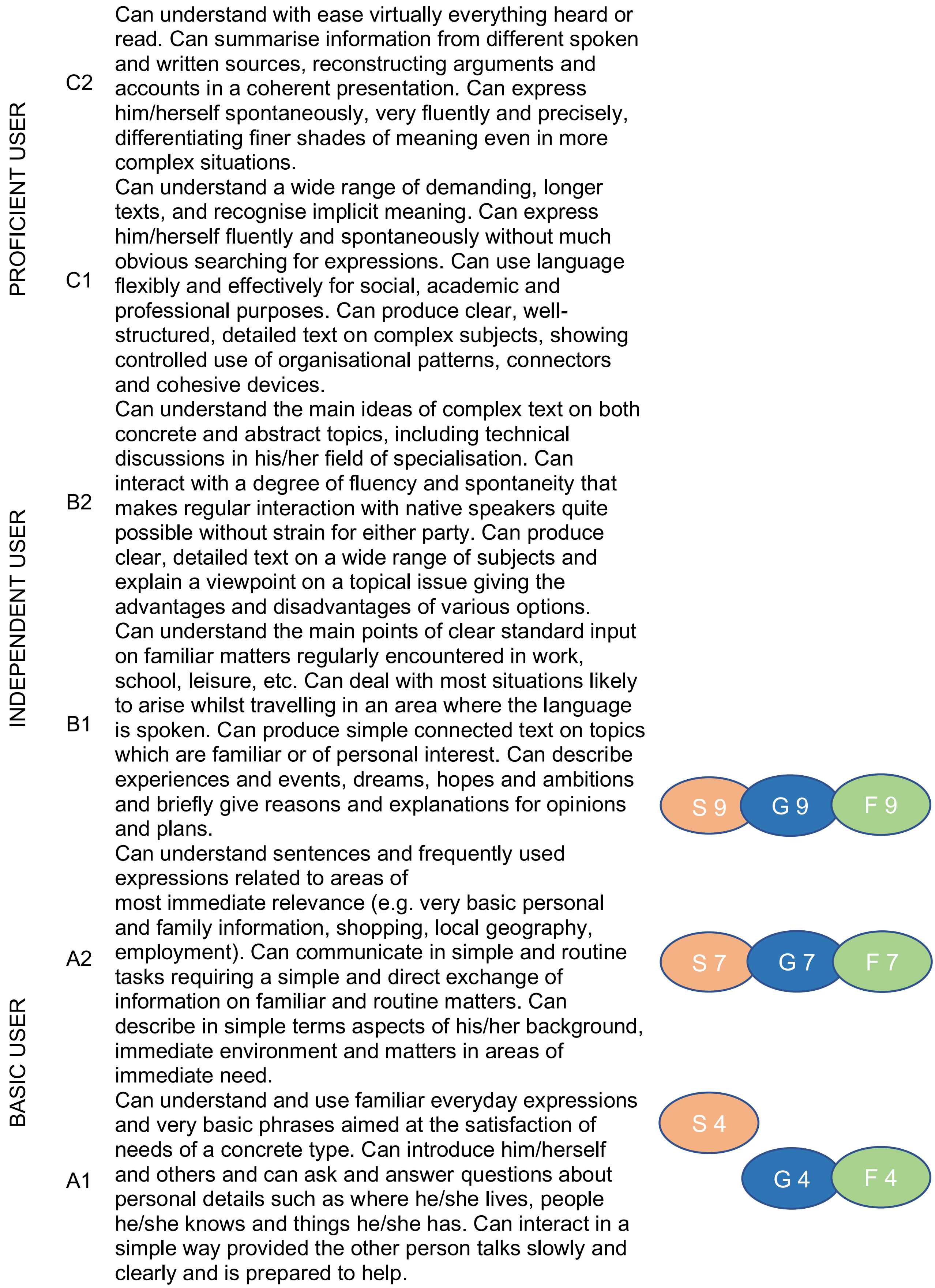
Table 46 below shows indicative linking at qualification level for each grade, based on averaging across the CEFR sub-levels of components. The sub-levels were transformed into the numerical scale from Table 9, and the numerical values averaged to derive overall sub-level and then level for each language and grade. This is additionally depicted in Figure 44 on the next page.

Table 46 *Indicative linking at qualification level*

| Language | Grade | **Average** | SD | Min | Max | Sub-level | Level |
|---|---|---|---|---|---|---|---|
| Spanish | 4 | **1.42** | 0.46 | 0.67 | 2.00 | A1 high | A1 |
| | 7 | **2.00** | 0.23 | 1.67 | 2.33 | A2 mid | A2 |
| | 9 | **2.92** | 0.22 | 2.67 | 3.33 | B1 low | B1 |
| German | 4 | **1.25** | 0.32 | 0.67 | 1.67 | A1 mid | A1 |
| | 7 | **2.25** | 0.22 | 2.00 | 2.67 | A2 mid | A2 |
| | 9 | **2.75** | 0.22 | 2.33 | 3.00 | B1 low | B1 |
| French | 4 | **1.17** | 0.37 | 0.67 | 1.67 | A1 mid | A1 |
| | 7 | **2.25** | 0.52 | 1.33 | 3.00 | A2 mid | A2 |
| | 9 | **2.83** | 0.29 | 2.33 | 3.33 | B1 low | B1 |

It appears that performance standards between the 3 languages are reasonably aligned at qualification level despite some component-level differences. The results suggest that grade 4 is around high A1 level for Spanish and mid A1 level for German and French. Grade 7 is around mid A2 level and grade 9 around low B1 level for all languages. This result accords with the results of the content mapping, which suggested that each of the 3 GCSE MFL specifications assessed most of the skills up to A2+ (i.e. high A2) level, with some aspects of language competence assessed up to low B1 level.

In addition to the limitations already discussed in the Limitations section, an important "health warning" regarding the interpretation of this linking is in order. It should be borne in mind that certain aspects of assessments noted in previous sections, and highlighted in both content mapping and in discussion with panellists, particularly with respect to assessment of interaction and integrated skills, would to some extent limit the interpretation based on these assessments that candidates are fully at A2 or B1 level. This is because the assessments themselves provide little evidence of some of the skills essential for communicative language competence, such as ability to engage in meaningful interaction. In a sense, it may be more appropriate to say that, overall, candidates achieving each of the GCSE grades possess most, but not all the skills and knowledge of the CEFR level assigned in this linking exercise. While this is also true of A2 level to some extent, most of the caveats and discrepancies noted above relate to where assessments appear to be targeting B1 level, as in many cases assessments were patchy in the extent to which they allowed for all of the skills relevant for B1 level to be demonstrated. This would mean that the levels assigned to different grades could be seen as overestimates to some extent, particularly for B1 level, but also to some extent for A2. This should be borne in mind in any discussions about whether A2 or B1 level may be appropriate for different GCSE grades.



PROFICIENT USER

C2 Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.

C1 Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.

INDEPENDENT USER

B2 Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

B1 Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.

BASIC USER

A2 Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.

A1 Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Figure 44 *Estimated qualification level mapping for each language and grade*

This linking study dealt with describing the content/construct of GCSE MFL specifications and tests, as well as performances, in terms of the CEFR, and relating the current GCSE grading standards to the CEFR. Therefore, this linking is not a statement of what the GCSE standard should be, but an approximate description of what the performance and assessment/grading standard currently appears to be, using the language and descriptors of the CEFR. The results essentially give an indication of where GCSE assessments are pitched and which performance standards are represented by different GCSE grades, using the language of the CEFR descriptors.

We have noted in several places previously that the GCSE MFL assessments reviewed in this study do not appear to elicit sufficient evidence of certain linguistic skills that may be considered by some to be a crucial part of communicative language competence. It would seem important to investigate these issues further and explore ways in which the assessments might be made more effective in assessing these important skills. We hope that this study has demonstrated that relating a conceptualisation of linguistic ability to the methods of assessment can be useful in highlighting both the desirable features of assessments in relation to their subject matter, and gaps in their ability to provide evidence of the relevant aspects of their subject matter. As far as GCSE MFLs should enable learners to act in real-life situations, expressing themselves and accomplishing tasks of different natures, it would make sense that, like the CEFR, they put the co-construction of meaning (through interaction) at the centre of both learning and assessment process.

The linking results are offered to stakeholders for consideration as to whether the content and performance standards and assessment demands associated with the key GCSE grades are appropriate given the purpose of GCSE qualifications, the spirit and nature of the curriculum, and the current context of GCSE MFL learning and teaching. For instance, if the relevant stakeholders were to conclude that, generally speaking, mid A2 level of performance is an appropriate expectation for GCSE grade 7 in terms of what learners can do, then this would mean that the current grading standard is in fact also appropriate (as long as the assessments do not include too many invalid sources of difficulty). If, on the other hand, the conclusion was that this level is too high for GCSE grade 7, this could provide rationale to support a change to grading standards. However, in this case, this rationale would not be based on statistical evidence or any notions of comparable ‘value-added’ between different subjects, but based on an understanding of what an appropriate performance standard, in terms of what students can do, is or should be for each grade within MFLs themselves.

We would suggest, however, in the spirit of the CEFR, that discussions around the appropriateness of language performance and assessment standards should consider important aspects of the context of language teaching in schools. CEFR (Council of Europe, 2018: 28) suggests planning backwards from learners’ real life communicative needs, with consequent alignment between curriculum, teaching and assessment. As North (2007a) points out, educational standards must always take account of the needs and abilities of the learners in the context concerned. Norms of performance need to be definitions of performance that can realistically be expected, rather than relating standards to “some neat and tidy intuitive ideal” (Clark 1987: 46, cited in North, 2007a). This posits an empirical basis to the definition of standards. If used appropriately, the CEFR could aid this endeavour in the context of GCSE MFLs in England.

# References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research & Perspectives*, *9 (1)*, 95–104.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement 2*: 449–60.

Asch, S.E. (1956). Studies of independence and conformity: a minority of one against a unanimous majority*. Psychological Monographs*, *70, 9, (whole no. 416),* 1–70. http://dx.doi.org/10.1037/h0093718

Baron, R. S. (2005). So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. In M. P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 37*, 219-253. San Diego, CA, US: Elsevier Academic Press. http://dx.doi.org/10.1016/S0065-2601(05)37004-3

Deutsch, M., & Gerard, H.B. (1955). A study of normative and informational influence upon individual judgement. *Journal of Abnormal and Social Psychology, 51*, 629–636. http://dx.doi.org/10.1037/h0046408

Bartram, B. (2005). Choice and the French curriculum – pupil views from around Europe. *Francophonie 32*: 3–6.

Bauckham, I. (2018). *A nation of monoglots? How we need to change the way we think about and learn languages in our schools*. Talk to the Conservative Education Society on Monday, 9th July 2018. Retrieved from https://johnbald.typepad.com/language/2018/07/ian-bauckham-cbe-what-has-gone-wrong-with-language-teaching-and-what-we-can-do-about-it-.html

Bauckham, I. (2016). *Modern foreign languages pedagogy review: A review of modern foreign languages teaching practice in key stage 3 and key stage 4.* Review report written for the Teaching Schools Council. Retrieved from https://www.tscouncil.org.uk/wp-content/uploads/2016/12/MFL-Pedagogy-Review-Report-2.pdf

Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations, *Research Papers in Education, 23:3*, 357-373. https://doi.org/10.1080/02671520701755440

Block, D. (2002). Communicative language teaching revisited: discourses in conflict and foreign national teachers. *Language Learning Journal 26*: 19–26. https://doi.org/10.1080/09571730285200191

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* 2nd ed. Mahwah, NJ: Lawrence Erlbaum.

Bramley, T. (2007*).* Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). London, U.K.: Qualifications and Curriculum Authority.

Brunfaut, T., & Harding, L. (2013). *Linking the GEPT Listening Test to the Common European Framework of Reference.* (LTTC-GEPT Research Report RG-05.) Retrieved from https://eprints.lancs.ac.uk/id/eprint/69811/1/Brunfaut_Harding2014_GEPT_listening_test_linking_study.pdf

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6(4),* 284–290.

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

Coe, R. (2008). Comparability of GCSE examinations in different subjects: An application of the Rasch method. *Oxford Review of Education*, *34 (5),* 609–36. https://doi.org/10.1080/03054980801970312

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Cambridge: Cambridge University Press. Retrieved from http://assets.cambridge.org/052180/3136/sample/0521803136ws.pdf

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors.* Council of Europe. Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Cuff, B. M. P., Meadows, M. & Black, B. (2018). An investigation into the Sawtooth Effect in secondary school assessments in England. *Assessment in Education: Principles, Policy & Practice*, 26(3), 321–339. https://doi.org/10.1080/0969594X.2018.1513907

Cuff, B. M. P. (2017). *Perceptions of subject difficulty and subject choices: Are the two linked, and if so, how?* (Report No. Ofqual/17/6288). Coventry, UK: Ofqual. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/649891/Perceptions_of_subject_difficulty_and_subject_choices.pdf

Curcin, M., & Black, B.  (in prep.) *Improving awarding: 2018/2019 pilots*. Unpublished Report. Coventry, UK: Ofqual.

Curcin, M., & Black, B. (2019). *Awarder and stakeholder surveys on GCSE MFL performance standards.* (Report No. Ofqual/19/6560). Coventry, UK: Ofqual.

Curcin, M., & Black, B. (2018) *Inter-subject Comparability: Higher Education representatives' perception of grade standard adjustment in some MFL and science A levels.* (Report No. Ofqual/18/6450/4). Coventry, UK: Ofqual.

Dearing, R., & King, L. (2007). *Languages review*. Nottingham, UK: Department for Education and Skills. Retrieved from https://www.languagescompany.com/wp-content/uploads/the-languages-review.pdf

De Jong, J. H. A. L. (2009). *Unwarranted claims about CEF alignment of some international English language tests.* Paper presented at EALTA, Turku, Finland. Retrieved from http://www.ealta.eu.org/conference/2009/docs/friday/John_deJong.pdf

Department for Education (DfE). (2014). *Modern Foreign Languages. GCSE subject content.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/485567/GCSE_subject_content_modern_foreign_langs.pdf

Department for Education (DfE). (2013). Statutory Guidance, National Curriculum in England: Languages Programmes of Study. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-languages-progammes-of-study/national-curriculum-in-england-languages-progammes-of-study

Department of Education and Science/ Welsh Office (DES/WO). (1990). *Modern Foreign Languages for Ages 11–16: Proposals of the Secretary of State for Education and Science and the Secretary of State for Wales.* Great Britain, Department of Education and Science.

Department for Education and Science/Welsh Office (DES/WO). (1991*). Modern Foreign Languages in the National Curriculum.* London, UK: HMSO.

Engelhard, G., Jr., Kobrin, J. L., &  and Wind, S. A. (2014). Exploring differential subgroup functioning on SAT writing items: What happens when English is not a test taker's best language? *International Journal of Testing*, *14*, 339–359. https://doi.org/10.1080/15305058.2014.931281

Figueras, N., North, B. (Dir), Takala, S., van Avermaet, P., & Verhelst, N. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR): A Manual, Strasbourg: Council of Europe. Retrieved from https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Graham, S. (2002). Experiences of learning French: a snapshot at years 11, 12 and 13. *Language Learning Journal 25 (1)*,15–20. https://doi.org/10.1080/09571730285200051

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology 8 (1)*, 23–34.

Jones, N. (2009). A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting. *Research Notes, 37*, 6–9. https://www.cambridgeenglish.org/images/23156-research-notes-37.pdf

Lockyer, C., & Newton, P. E. (2015). Inter-subject comparability: a review of the technical literature. (Report No. Ofqual/15/5794). Coventry, UK: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/606043/2-inter-subject-comparability-a-review-of-the-technical-literature.pdf

Kaftandjieva, F. (2009). *Basket procedure: The breadbasket or the basket case of standard setting methods?* In Figueras, N. & J. Noijons (eds.) *Linking to the CEFR levels: Research perspectives*. Arnhem, The Netherlands: Cito, EALTA. http://www.ealta.eu.org/documents/resources/Research_Colloquium_report.pdf

Kaftandjieva, F. (2010). *Methods for Setting Cut Scores in Criterion referenced Achievement Tests: A comparative analysis of six recent methods with an application to tests of reading in EFL.* Arnhem, The Netherlands: Cito, ELTA. http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, *1 (2),* 152–176.

Klapper, J. (1997). Language learning at school and university: the great grammar debate continues (I). *Language Learning Journal 16*, 22–27. https://doi.org/10.1080/09571739785200231

Klapper, J. (1998). Language learning at school and university: the great grammar debate continues (II). *Language Learning Journal 18*, 22–28. https://doi.org/10.1080/09571739885200211

Klapper, J. (2003). Taking communication to task? A critical review of recent trends in language teaching. *Language Learning Journal 27*, 33–42. https://doi.org/10.1080/09571730385200061

Linacre, J. M. (2011). *A user's guide to FACETS Rasch-model computer programs. Program Manual 3.68.1. Chicago, IL: Winsteps.com.*

Long, M. (1991). Focus on form: a design feature in language teaching methodology. In *Foreign language research in crosscultural perspective, eds. K. de Bot, R.B. Ginsberg and C. Kramsch*, 39–52. Amsterdam: Benjamins. http://doi.org/10.1075/sibil.2.07lon

Macaro, E. (2008). The decline in language learning in England: getting the facts right and getting real. *Language Learning Journal 36, no. 1*:,101–108. https://doi.org/10.1080/09571730801988595

Martyniuk, W. (Ed.) (2010). *Aligning Tests with the CEFR. Reflections on using the Council of Europe's draft Manual*. Cambridge: Cambridge University Press.

Meiring, L., & N. Norman. (2001). Grammar in MFL teaching revisited. *Language Learning Journal 23, no. 1*, 58–66. https://doi.org/10.1080/09571730185200101

Mitchell, R. (2000). Anniversary article. Applied linguistics and evidence-based classroom practice: the case of foreign language grammar pedagogy. *Applied Linguistics 21, no. 3*, 281–303. https://doi.org/10.1093/applin/21.3.281

Mitchell, R. (1994). The communicative approach to language teaching: an introduction. In A. Swarbrick (Ed) *Teaching Modern Languages (pp.33–42)*. London: Routledge.

Mitchell, R., & Martin, C. (1997). Rote learning, creativity and understanding in the foreign language classroom. *Language Teaching Research 1, no. 1*, 1–27. https://doi.org/10.1177/136216889700100102

Mitchener, G. W. (2016). Spontaneous Language. *Encyclopedia of Evolutionary Psychological Science*, 1–5.

Morgan-Short, K., Steinhaurer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and Implicit Second Language Training Differentially Affect the Achievement of Native-like Brain Activation Patterns. *Journal of Cognitive Neuroscience, 24 (4)*, 933–947. https://doi.org/10.1162/jocn_a_00119

Moscovici, S., & Zavalloni, M. (1969). The group as the polarizer of attitudes. *Journal of Personality and Social Psychology, 12*, 125–135. https://doi.org/10.1037/h0027568

Newton, P. E. (2012). Making sense of decades of debate on inter-subject comparability in England. *Assessment in Education: Principles, Policy & Practice*, *19 (2),* 251–273. https://doi.org/10.1080/0969594X.2011.563357

North, B. (2007a, February). *The CEFR Common Reference Levels: validated reference points and local strategies*. Presented at The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities, Strasbourg.

North, B. (2007b). The CEFR Illustrative Descriptor Scales. *The Modern Language Journal, 91(4)*, 656-659. https://doi.org/10.1111/j.1540-4781.2007.00627_3.x

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15 (2)*, 217–263. https://doi.org/10.1177/026553229801500204

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Nunan, D. (1989). *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press.

O'Sullivan, B. (2015). *Linking the Aptis Reporting Scales to the CEFR*. (Report No. TR/2015/003). UK: British Council. https://www.britishcouncil.org/sites/default/files/tech_003_barry_osullivan_linking_aptis_v4_single_pages_0.pdf

Pachler, N. (2000). Re-examining communicative language teaching. In K. Field (Ed), *Issues in Modern Foreign Language Teaching* (pp. 22–37). London: Routledge Falmer.

Pollitt, A, Ahmed, A, Baird, J-A, Tognolini, J, and Davidson, M. (2008). *Improving the quality of GCSE Assessment*. A report for Qualifications and Curriculum Authority. Downloaded on 01/09/14 from: http://www2.ofqual.gov.uk/downloads/category/106-gq-monitoring?download=352%3Aimproving-the-quality-of-gcse-assessment-january-2008

Raikes, N., Scorey, S., & Shiell, H. (2008, September). *Grading examinations using expert judgements from a diverse pool of judges*. A paper presented to the 34th annual conference of the International Association for Educational Assessment, Cambridge.

Savignon, Sandra J. (2000). Communicative language teaching. In M. Byram, M (Ed.), *Routledge Encyclopedia of Language Teaching and Learning (pp.124–129)*. London: Routledge

Smith, R. M., Schumacker, R. E., & Bush, J. J. (1998). Examining replication effects in Raschfit statistics. In M. Wilson & G. Engelhard, Jr. (Eds.), *Objective Measurement: Theory into practice vol. 5,* (pp. 303–317). Stanford, CT: Ablex Publishing Corp.

Stratton, T., & Zanini, N. (2019). *Evaluating the impact of the introduction of new GCSE MFL assessments in 2018*. (Report No. Ofqual/19/6561). Coventry, UK: Ofqual.

Swan, M. (1985). A Critical look at the Communicative Approach (2). *ELT Journal, 39 (2),* 76–87 http://seas3.elte.hu/coursematerial/HalapiMagdolna/Swan2.pdf

Taylor, L. (2004). Issues of test comparability. *Research Notes, 15*, 2–5. https://www.cambridgeenglish.org/Images/23131-research-notes-15.pdf

The Quality Assurance Agency for Higher Education (2015). Subject Benchmark Statement: Languages, Cultures and Societies. Retrieved from https://www.qaa.ac.uk/docs/qaa/subject-benchmark-statements/sbs-languages-cultures-and-societies-15.pdf?sfvrsn=2098f781_12

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review 3*, 273–86.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement, *Assessment in Education: Principles, Policy & Practice.* https://doi.org/10.1080/0969594X.2019.1602027

Wingate, U. (2018). Lots of games and little challenge – a snapshot of modern foreign language teaching in English secondary schools. *The Language Learning Journal, 46 (4)*, 442–455. https://doi.org/10.1080/09571736.2016.1161061

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8:3*, 370–371. https://rasch.org/rmt/rmt83b.htm

Wu, J. R. W., & Wu, R. Y. F. (2010). Relating the GEPT Reading Comprehension Test to the CEFR. In Waldemar Martyniuk (Ed.), *Aligning Tests with the CEFR: Case studies and reflections on the use of the Council of Europe's Draft Manual* (pp. 204–224). Cambridge: Cambridge University Press

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

**October 2019** **Ofqual/19/6559/1**