RESEARCH AND ANALYSIS

Understanding ESOL Skills for Life qualifications

Assessment properties, comparability, core curriculum and CEFR alignment

ofqual

Authors

- Milja Curcin
- Stuart Cadwallader
- Ezekiel Sweiry

Acknowledgements

We would like to thank the many people without whose input 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, discussions and analytical support, and various ad hoc and last-minute requests for help)

Contents

List of acronyms

AO – Awarding Organisation

AC – Assessment criteria

CBA – Competence-based assessment

CEFR – Common European Framework of Reference for Languages

CJ – Comparative judgement

EAL – English as an Additional Language

EFL – English as a Foreign Language

ESFA – Education and Skills Funding Agency

ESL – English as a Second Language

ESOL – English for Speakers of Other Languages

FSE – Functional Skills English

GCR – General Conditions of Recognition

GLH – Guided learning hours

IELTS – International English Language Testing System

LO – Learning outcome

MFL – Modern Foreign Language

NARIC – National Academic Recognition Information Centre

NSAL – National Standards for Adult Literacy

NQF – National Qualifications Framework

NVQ – National Vocational Qualification

SFL – Skills for Life

QCA – Qualifications and Curriculum Authority

QCF – Qualifications and Credit Framework

RQF – Regulated Qualifications Framework

SELT – Secure English Language Test

VTQs – Vocational, technical and other general qualifications

List of tables

List of figures

Executive summary

Background

In the corporate plan 2019-2022, Ofqual committed to considering the effectiveness of the regulatory requirements relating to qualifications covering English for speakers of other languages (ESOL).

This study focuses on the ESOL Skills for Life (SfL) qualifications.[1] These were developed as part of a suite of basic skills qualifications to help migrants resident in this country to integrate better with English-speaking communities and to access education and employment, helping them to fulfil their potential.

The SfL qualifications were accredited by Ofqual in 2014, against the 2011 ESOL criteria and the rules of the Qualifications and Credit Framework (QCF). They must meet the requirements of the National Standards for Adult Literacy (NSAL) (QCA, 2000) and demonstrate a clear relationship to the Adult ESOL Core Curriculum (DfES, 2001). They are government-funded qualifications, offered at levels Entry 1, Entry 2, Entry 3, Level 1 and Level 2 of the Regulated Qualifications Framework (RQF), delivered by several awarding organisations (AOs).

The SfL qualifications operate in a challenging context of language assessment for migration and integration. The challenge is the sheer variety of the experience of the target cohort for these qualifications in terms of literacy levels, language and educational backgrounds, life histories including trauma and other vulnerabilities, aspirations and needs within the UK society. This often calls for flexibility in delivery and assessment approaches. On the other hand, employers, education providers and other users may rely on the results of these qualifications and will need to know there is sufficient comparability between different awarding organisations (AOs).

Study aims and research questions

The purpose of this research was to gain an understanding of how effective our current regulatory requirements are for these qualifications in making sure they function as intended in relation to their purposes and to inform decisions on our

[1] The other type of ESOL qualifications on Ofqual's register is ESOL International qualifications. These are suitable for primarily over-16 learners whose first language is not English and who are living abroad. They are recognised for entry to higher education in the UK, while some have been recognised as Secure English Language Tests (SELTS), which can be used to satisfy visa, residency and citizenship requirements. These qualifications reference the Common European Framework of Reference for Languages (CEFR) and have a minimum of 75% external assessment.

requirements in future. We explored and evaluated a sample of SfL qualifications in terms of their design and assessment properties, alignment with the core curriculum, inter-AO comparability and validity. Specifically, we wanted to develop a clearer understanding of:

- assessment models of SfL qualifications
- how AOs approach the development and delivery of SfL qualifications given their understanding of the cohort needs, qualification purposes and the curriculum
- how different SfL qualifications align with the core curriculum
- how different SfL qualifications align with each other as well as with the Common European Framework of Reference for Languages (CEFR) in terms of assessment and performance standards

The research involved a sample of 4 AOs, which account for a large proportion of the certifications in ESOL SfL, with 2 additional AOs included in interviews. Due to resource and time limitations, we focused on a subset of qualification levels. The research was conducted in several strands:

- Study 1: Assessment review (focus on Entry 1 and Entry 3)
- Study 2: Interviews with key members of AO staff
- Study 3: Curriculum alignment review (focus on Entry 1, Entry 3 and Level 1)
- Study 4: Comparability study involving content mapping and linking of pass grade standards to the CEFR (focus on Entry 1, Entry 3 and Level 1)

Key findings

This report brings together the findings from our 4 studies and suggest some potential areas for improvement. This section is organised according to the key themes that emerged across different studies or from our background research.

Issues of recognition

It is clear from the views expressed in our interviews with AO staff, as well as from the literature, that the SfL qualifications are seen by practitioners as having enormous value to learners and society as a whole. The AO staff saw the SfL qualifications as similar in their enabling function to ESOL International qualifications. Some of these are widely recognised, at competence levels similar to Entry 3, Level 1 or Level 2, by reputable higher education institutions or for job recruitment.

In contrast, our interview participants felt that the potential of SfL qualifications to enable learner progression was not sufficiently recognised. In our interviews, as well

as in the relevant literature, there was also a sense that these qualifications are undervalued in terms of funding. The AOs told us that this affected user attitudes towards the qualifications, including centres, employers, higher education institutions, and learners themselves.

Multiple qualification purposes

The views expressed in our interviews as well as in the literature suggest that the SfL qualifications potentially have multiple purposes and there is insufficient clarity about their relative importance to different users (for instance, learners, centres or employers). Their purposes are currently not clearly articulated in Ofqual ESOL criteria.

Using Newton's (2017) perspectives framework for thinking about assessment purposes, some of SfL qualification purposes can be seen from an **engagement perspective,** helping to ensure that learning actually occurs, inasmuch as they are seen as valuable for (linguistic) integration of migrants and for removing barriers to accessing employment or education. Given their required alignment with the core curriculum, it appears that some of the purposes relate to the **expertise perspective**, providing evidence of a level of mastery of the specified content domain, as well as boosting confidence and motivation to build on skills and expand the expertise within the domain. The SfL qualifications are seen as both socially and personally valuable in relation to these 2 perspectives. What is less clear, based on the issues with recognition of these qualifications for progression to employment or education, is how far they have (or could have) purposes related to **information perspective** that are valuable to stakeholders beyond learners or qualification providers as the basis for making any decisions.

This situation may indirectly impact on overall quality and appropriateness of provision and reduce the extent to which any of the intended purposes are truly fulfilled, potentially reducing the usefulness of these qualifications to users. It is important to seek clarity about the intended purposes of these qualifications and their relative importance. This is necessary to enable coherent considerations of the most appropriate assessment models, as well as approaches to regulation, and any trade-offs.

Assessment models

A wide range of assessment models and approaches are used across the 4 AOs in our sample. Some of the differences in approaches appear to stem from a desire to offer flexibility of choice to learners and centres. The AOs in our interviews considered the flexibility of the SfL qualification structure, its modularity and the variety of assessment models (such as an option to use internal or external

assessment at all levels) as beneficial and justified by the wide range of learner and centre contexts.

Our assessment review, however, highlighted quality issues with some of the assessment approaches and assessment instruments. In some cases, the assessments were not entirely appropriate for the construct of assessment or were poorly executed in practice. The AO interviews suggested that some of the assessment design choices were historical and not always reflective of best practice in language assessment but AOs were not always able to provide a clear rationale for them beyond citing original requirements that had not been updated.

Curriculum alignment

Our reviewers thought that the curriculum was largely fit for purpose, sufficiently clear and indicative of the appropriate qualification levels. However, the review suggests that the SfL qualifications may not be sufficiently well aligned with the curriculum. This was primarily in terms of demand, aspects of coverage (for instance, narrowing of the curriculum) and interpretation of the curriculum messages.

At Entry 3 and Level 1, the demand of assessments and pass grade standards were deemed to be too low in most cases except in speaking. Entry 1 was seen as broadly appropriate, if occasionally too high in relation to the curriculum standards. The overall quality and appropriateness of coverage and relative demand of the assessments seem to decline from Entry 1 to Level 1 for most of the AOs in our sample, with insufficient stretch at higher levels, suggesting threats to appropriate candidate progression.

CEFR alignment

The CEFR describes second language ability in terms of a ‘can do’ scale that reflects increasing complexity and range of language competence from basic user (A1, A2), independent user (B1, B2) to proficient user (C1, C2) (see Appendix 1 for the CEFR global scale descriptors). The indicative CEFR content mapping produced in our studies suggests there is meaningful alignment of the SfL qualifications with the CEFR in many respects, and thus also with ESOL International qualifications.

The CEFR content mapping placed Entry 1 reading and writing components at A1/A1+ level. Entry 1 speaking was judged to be at A1+/A2 level. For writing, the standard linking exercise additionally indicated that the pass grade standard is at about A1+ level. However, there is a separate question as to whether A1+ level or higher is suitable for an initial Entry level qualification (even if it is deemed appropriate according to the curriculum). Given that Entry 1 pass grade standard appears to be significantly higher than threshold A1 CEFR level at least in writing,

this may suggest a potential need for separate qualifications at somewhat lower level ('pre-entry') for learners who are starting from the point of no English at Entry 1.

Entry 3 components were judged in the content mapping to be at B1/B1+ level. However, the results of our standard linking study for reading comprehension suggest that actual pass grade standards of some exams are lower, at A2+, and in all cases closer to B1 threshold level than B1+. Similarly, the standard linking for the writing components indicated that the pass grade performance standard at Entry 3 is A2+/low B1 at most, while Level 1 was mapped to B1+/low B2.

Perhaps the most significant finding from the CEFR standard linking study for writing is that there does not appear to be enough differentiation between higher levels of the SfL qualifications at the pass grade threshold, particularly between Entry 2 and Entry 3. Both of these levels were mapped onto A2+ CEFR level for some qualifications. This is also the case to some extent between Entry 3 and Level 1 writing qualifications. Thus, the findings from the CEFR standard linking largely confirm the findings from the curriculum review, namely, that Entry 3 and Level 1 SfL qualifications may not be sufficiently demanding given the demand implied in the curriculum and relative to the demand of the lower levels. This lack of stretch at Entry 3 in particular may be contributing to the perceived "large jump" in standards at Level 1, which was noted by our curriculum reviewers.

The CEFR content mapping highlighted the discrepancies with the CEFR in relation to assessment of literacy in SfL qualifications, for example, assessment of Latin alphabet knowledge, emphasis on orthographic accuracy at Entry 1 or emphasis on awareness of the writing process at Entry 3, which could be seen as too advanced for learners at those levels, particularly in the migration context. Such requirements in the domain of literacy in SfL qualifications are likely related to their link to the NSAL, which are primarily targeted at learners for whom English is their first language. This potentially raises questions about whether mapping qualifications such as SfL, which is an English as a second language qualification, to standards such as NSAL, intended for English as the first language qualifications (for instance, Functional Skills) is entirely appropriate.

Comparability of pass grade standards

Pass grade standards in Entry 3 reading comprehension assessments were broadly comparable and at around threshold B1 level for 3 out of 4 AOs. For one AO, the pass grade standard was mapped to A2+, and is thus lower than the rest. This is despite the overall profile and demand of the tests from the 4 AOs being reasonably aligned.

Similarly, some comparability issues were noted in our CEFR linking study of pass grade standards in writing, at Entry 1 level. All 4 AOs, however, were mapped to A1

level on the CEFR, despite some discrepancies at the pass grade within this level. These comparability issues, therefore, are of somewhat less concern than those identified in reading comprehension.

Potential improvements

In light of these findings, Ofqual (and other relevant stakeholders where appropriate, for instance, the DfE) should consider how the qualifications could be improved to address some of the issues identified. Some areas for improvement are listed below:

1. The range, relative priority, and interactions of SfL qualification purposes should be clarified. Given the complexity of SfL learner needs and the importance of engagement-related purposes of these qualifications, revisiting the trade-off between these and any information-related purposes seems particularly important as a starting point. This should help to inform the choice of the most appropriate assessment models and any trade-offs between the extent of standardisation and flexibility in assessment design and delivery.
2. There are quality and validity issues in some individual approaches and assessment instruments, which need to be addressed. In addition, appropriateness of approaches that may be historically motivated rather than meaningfully linked to qualification purposes and constructs should be reconsidered. Assessment approaches should be sufficiently justified in terms of validity and best assessment practice.
3. The way the curriculum is covered and interpreted in some assessments may need to be improved to reflect its intention and implied demand.
4. Insufficient comparability between Entry 3 reading comprehension pass grade standards should be addressed if comparability were deemed important for certain purposes of SfL qualifications. Wider checks on comparability (between other components, levels and AOs) may also be warranted depending on the extent of comparability required.
5. Differentiation and progression between SfL qualification levels should be improved to ensure optimal progression for learners.
6. Mapping SfL qualifications to the CEFR might clarify to users how the skills at each RQF level relate to the skills of ESOL International learners, who might be competing for the same jobs or education courses upon entry to this country, and improve recognition of SfL qualifications for similar purposes.

Introduction

In the corporate plan 2019-2022 (Ofqual, 2019), Ofqual committed to considering the effectiveness of the regulatory requirements relating to qualifications covering English for speakers of other languages (ESOL)[2]. This was originally in response to the government's English Language Strategy (Integrated Communities Strategy green paper, 2019). This work is now feeding into the DfE review of Level 2 and below qualifications.

There are 2 types of ESOL qualifications currently on the Ofqual register:

- **ESOL Skills for Life (SfL) qualifications.** These are suitable for primarily over-16 learners whose first language is not English, who are resident in the UK, and wish to increase their English language knowledge and proficiency for life, work or further study. They were accredited by Ofqual in 2014, against the 2011 ESOL criteria and the rules of the Qualifications and Credit Framework (QCF). These qualifications must meet the requirements of the National Standards for Adult Literacy (henceforth, NSAL) (QCA, 2000) and demonstrate a clear relationship to the Adult ESOL Core Curriculum (DfES, 2001). They are government-funded qualifications, offered at levels Entry 1, Entry 2, Entry 3, Level 1 and Level 2 of the Regulated Qualifications Framework (RQF).
- **ESOL International qualifications.** These are suitable for primarily over-16 learners whose first language is not English and who are living abroad. These qualifications are recognised for entry to higher education in the UK. Some ESOL International qualifications have been recognised as Secure English Language Tests (SELTS), which can be used to satisfy visa, residency and citizenship requirements. ESOL International qualifications are primarily delivered abroad, but also within the UK. These qualifications reference the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001)[3] as well as the RQF levels and have a minimum of 75% external assessment.

[2] Acronyms other than ESOL are widely used in English language teaching, including English as an Additional Language (EAL), English as a Second Language (ESL) and English as a Foreign Language (EFL). In post-16 education the universally accepted term is ESOL, whereas EAL is used to describe this work in schools (DfES, 2001). EFL is commonly used to denote ESOL International qualifications.

[3] The CEFR is an internationally widely used framework describing language ability via a common 'can do' scale, allowing broad comparisons across languages and qualifications.

The focus of the current research was on ESOL SfL qualifications. This focus was partly motivated by the fact that these qualifications are government funded, as well as by the fact that Ofqual criteria for them are fairly minimal.

The purpose of this research was to explore and evaluate a sample of SfL qualifications in terms of their design and assessment properties, alignment with current regulations/criteria, comparability between awarding organisations (AOs) and validity. We also wanted to learn about the broader context in which these qualifications operate. This would provide us with an understanding of the effectiveness of our current regulatory requirements for these qualifications in ensuring their optimal functioning and validity in relation to their purposes. This evidence base will contribute to policy development and decisions about future regulatory arrangements for SfL qualifications.

More specifically, we wanted to develop a clearer understanding of:

- AOs' assessment models for these qualifications
- how AOs approach the development and delivery of SfL qualifications given their understanding of the cohort needs, qualification purposes and the curriculum
- how different SfL qualifications align with the core curriculum
- how different SfL qualifications compare to each other as well as to the CEFR in terms of assessment and performance standards

The research was carried out in several strands:

- Study 1: Assessment strategy and design review (focus on Entry 1 and Entry 3)
- Study 2: Interviews with key members of AO staff
- Study 3: Curriculum alignment review (focus on Entry 1, Entry 3 and Level 1)
- Study 4: Comparability study involving
    - CEFR content mapping (focus on Entry 1 and Entry 3)
    - CEFR linking of assessment standards of reading comprehension tests (focus on Entry 3)
    - CEFR linking of writing performance samples (focus on Entry 1, Entry 3 and Level 1)

The research involved a sample of 4 AOs that participated across all of the studies, with 2 additional AOs included for the interviews. In addition, due to resource and time limitations, we focused on particular levels as exemplars, on the assumption that practices would be broadly similar across all levels within one AO.

It is important to note that, although this research identifies qualification properties and practices related to specific, regulated qualifications, it is not concerned with issues of compliance. The findings are examined through the general lens of validity and comparability, and in the context of broader issues in language assessment for migration and integration, rather than in relation to specific regulatory requirements. The research is intended to explore general practices and approaches used for development and delivery of SfL qualifications so that we could learn from this and ensure that improvements are made overall where issues are observed, and that our regulatory approach is such that it can ensure optimal functioning, appropriate standards, and validity of these qualifications.

The report is structured as follows. After presenting some background information about the SfL qualifications, we report on the details of each of the 4 studies in turn. The sections for each study present the relevant methodology and results. This is followed by a general discussion and recommendations section.

SfL qualifications – background information

In this section, we briefly outline some information about the SfL qualifications' purposes and design, cohort, funding, progression patterns and apparent issues with recognition of these qualifications by different stakeholders. We also discuss the implications of the apparent lack of clarity in SfL qualification purposes. This provides a backdrop and context to the findings presented in the rest of the report.

Qualification purposes and design

The SfL qualifications are government-funded qualifications that are currently used to foster and improve linguistic integration of different categories of migrants to the UK. They were previously used to evidence language skills for immigration purposes (for visa, residency and citizenship requirements), but this is no longer the case as this role has been taken over by the SELTs.

In one sense, the SfL qualifications and tests can be considered relatively low stakes, as their results are not used for decisions that directly affect learners' legal rights (for instance, their right to acquire citizenship). On the other hand, they are still intended and funded to serve an important function in relation to integration and promoting the chances of migrants to achieve their full potential within the UK society by removing or reducing language and literacy barriers (DfES, 2001; QCA, 2009). They help migrants already in this country to acquire language skills that might subsequently enable them to achieve the SELTs, perhaps for citizenship

requirements, or enable them to access employment or education courses. In this sense, they can still be seen as reasonably high-stakes qualifications, at least for some of their users.

The SfL qualifications were last accredited by Ofqual in 2014, against the 2011 ESOL criteria and the rules of the Qualifications and Credit Framework (QCF). Although the QCF was withdrawn in 2015 and the qualifications now have to abide by Ofqual General Conditions of Recognition these qualifications still currently largely follow, with some exceptions, the original QCF design rules.

The qualifications involve a full certificate in ESOL SfL and separate awards in reading, writing, and speaking and listening. Most AOs use the same QCF units, with 270 GLH (guided learning hours) across 3 units per level, while some have somewhat different units and GLH. The qualifications are offered at RQF levels Entry 1, Entry 2, Entry 3, Level 1 and Level 2. Individual awards can be taken at different levels, depending on candidates' skills profile and needs.[4] The certificate is awarded where all units are achieved at the level of the qualification. The qualifications from different AOs should be comparable in standards and are intended to be used interchangeably by learners within and between levels.

The key regulatory requirement for SfL qualifications is that they should meet the requirements of the NSAL and demonstrate a clear relationship to the core curriculum. Unlike ESOL International qualifications, the SfL qualifications are not required to be mapped to the CEFR.

The core curriculum itself is based on the NSAL. According to DfES (2001), the publication of the core curriculum followed national consultation with teachers and managers of ESOL programmes and relevant national bodies. It drew on a wide range of existing curricula from a variety of contexts, including the core curriculum for adult literacy, the National Literacy Strategy in schools and curricula for English as a Foreign Language (EFL). It is intended as a framework for English language learning and defines the skills, knowledge and understanding that speakers of other languages need in order to demonstrate achievement of the national standards in English, providing a reference tool for teachers of ESOL.

Within these parameters, awarding organisations have flexibility in:

---

[4] This aims to deal with the fact that most learners do not (need to) achieve the same language level in every skill (reading, listening, writing and speaking), typically exhibiting differentiated profiles of ability. It is common for learners to perform better (at a higher level) in the receptive skills (reading and listening) than in the productive skills (writing and speaking). Low-literate learners often perform better in the oral modes (listening and speaking) than in the written modes (reading and writing) (Council of Europe, 2020; ALTE, 2016). Similarly, different profiles might be appropriate for different purposes or needs. Such assessment of skills individually is considered good practice in the migration and integration context, though it may lead to negative impacts of certain areas of skills being neglected (ALTE, 2016).

- mode of assessment, including use of internal or external assessment in all skills areas and at all levels
- assessment approaches (such as task type, nature of mark schemes and approaches to grading)
- availability of assessments
- approach to resits
- whether to allow the use of dictionaries for reading assessments

Cohort properties

This level of flexibility in qualification design appears to relate to some extent to the nature of the SfL cohort, which is very diverse, with potentially a wide range of needs and goals in relation to English language learning. According to DfES (ibid., cf. DfES, 2000), the broad categories of learners who SfL qualifications are aimed at are:

- settled communities, including communities from the Asian sub-continent and Hong Kong. Some of these learners work long and irregular hours and cannot attend classes regularly
- refugees, who sub-divide into:
    - asylum seekers, most of whom are very keen to learn despite the challenges of resettlement and the trauma resulting from their recent experiences
    - settled refugees, many of whom have had professional jobs in the past, though some may have suffered a disrupted education due to war and unrest
- migrant workers, mostly from Europe, who are here to work and settle for most or all of their lives
- partners and spouses of migrants from all parts of the world, who are settled for a number of years and need to participate in the local community but are prevented by family responsibilities or low income from attending intensive EFL courses

Alongside their lack of familiarity with the cultural context of the UK, these learners have ‘ESOL needs’ in that they are learning English as an additional language, and their English needs further development, particularly with respect to writing, grammar and vocabulary (Roden & Osmaston, 2021) to enable their effective linguistic integration into the society. In this sense, SfL qualifications, as well as other qualifications for learners who are learning English as an additional language (for instance, ESOL International), have the broad purpose of enabling these learners to

develop their English skills and meet their ESOL needs. This fundamentally differentiates them from qualifications such as Functional Skills English (FSE) or GCSE English, which are designed for learners whose first language is English.

Beyond the ESOL needs, DfES (2001; cf. also QCA, 2009; ALTE, 2016) emphasises that within the above-mentioned learner groups the specific needs of learners vary considerably depending on their aspirations, educational background, language and literacy background and aptitude for learning languages. Learners' educational and employment backgrounds are often highly diverse, sometimes even within one teaching group. They may span a wide continuum, ranging from people with no previous education or employment at one end to highly educated professionals, such as doctors or university lecturers, at the other. Many learners already speak and write several languages and can use this knowledge to support their learning of English. Unlike many adult literacy learners for whom English is the first language, they often perceive themselves as successful learners and are very keen to learn. Some ESOL learners have left their countries unwillingly. Some are suffering from culture shock and are experiencing practical difficulties in specific areas of their lives, or racist attitudes and behaviour. Learners are typically settling in the UK and want to learn about the country's systems, such as education, how to apply for jobs, and rules concerning immigration.

More recently, Higton et al. (2019) found the following profile of learners in their ESOL provider survey sample: job seekers, people already in work, refugees, women not seeking work, recent migrants and asylum seekers and people of retirement age. Their survey findings showed that refugees were more likely to access pre-entry and Entry level provision and learners who were already in employment or receiving Jobseeker's Allowance were more likely to access learning at Level 1. They also noted that the profiles/demographics of those accessing English language learning provision changed during the past 10 to 15 years, resulting in the offer of more pre--entry and Entry level courses, to meet the increasing demand from learners. The changes include more asylum seekers and refugees wanting to learn English. Some of the providers they surveyed said they noticed a decrease of economic migrants from Eastern European countries while learners from Italy and Spain had increased. In general, however, it appears that the groups described in DfES (2001), for whom the curriculum and the SfL qualifications had been intended, are largely the same nowadays despite some shifts in the relative proportion of learners from different groups (see also Kings & Casey (2014) and WMCA (2019) for similar findings).

Funding, progression, and recognition issues

The SfL qualifications are publicly funded through a range of funding streams. They were fully funded for all learners until 2007; since then, the funding has gradually

decreased (Roden & Osmaston, ibid.; Kings & Casey, ibid.; Foster & Bolton, 2018). Currently, learners part pay for the course and assessment unless they are 16 to 19 years old, unemployed or on certain benefits, in which case the qualifications are fully funded. There is no funding for ESOL provided in the workplace (Foster & Bolton, ibid.).

Most certificates are issued at Entry level.[5] Furthermore, many of those learners who want to study at Level 1 appear to progress to other qualifications such as Level 1 FSE, either because higher-level ESOL provision is not available, or because they wish to study a mainstream qualification.[6] This trend is confirmed by Learning and Work Institute (2020) and Roden & Osmaston (ibid.). Roden & Osmaston observe that, even though the majority of the providers they surveyed indicated that most of their learners do progress to ESOL SfL Level 1 from Entry 3, a substantial proportion of learners progress to FSE and GCSE English qualifications. They also observed that this was more common for young people aged 16 to 19. The authors note that this may disadvantage these learners compared to the adults because the FSE/GCSE provision may not be suitable for ESOL learner needs as they are designed for learners whose first language is English and who are familiar with the cultural context of the UK.

Roden & Osmaston (ibid.) suggest the following interrelated reasons for some ESOL learners' progression onto FSE/GCSE English from Entry 3 SfL qualifications despite the former likely having too few guided learning hours for learners with ESOL needs, or being otherwise unsuitable for ESOL learners:

- government policy in favour of GCSE and FSE for 16 to 19 study programmes and apprenticeships has led some providers to insist that all young people should take these qualifications, regardless of language need
- adult providers may choose to offer FSE rather than ESOL as it is fully funded at higher levels, and they know that many learners are not able to pay for ESOL SfL courses
- learners and providers may choose FSE as they believe it is better recognised for future study and employment purposes

In study after study reviewed, spanning more than a decade (QCA, 2009; Kings & Casey, ibid.; the Casey review, 2016; Bell, Plumb & Marangozov, 2017; Foster & Bolton, ibid.; WMCA, ibid.; Higton et al., ibid.; Roden & Osmaston, ibid.; Rolfe &

[5] Note that the national data is distorted by the fact that Entry level is treated as one level, and many ESOL learners take several years to complete Entry level (Roden & Osmaston, ibid., cf. Kings & Casey, 2014). Therefore, the more appropriate comparisons between numbers of enrolments at entry vs. higher levels should be between Entry 3 and Level 1, rather than the whole of Entry vs. Level 1.

[6] Feedback from awarding organisations at an Ofqual ESOL workshop on 8 July 2019.

Stevenson, 2021), there are testimonies from qualification providers and other stakeholders about the high demand and need for specialised ESOL provision, as well as calls to improve the funding for all ESOL levels including pre-entry[7] and to increase the recognition of the SfL qualifications for progression to employment or further study. There are also calls to introduce funded Level 3 ESOL qualifications (or fund existing ESOL International qualifications such as IELTS) to enable learner progression to more specialised, high value-added jobs or higher education.

In this sense, ESOL SfL qualifications appear to stand in contrast with a number of ESOL International qualifications, which are also designed for learners with ESOL needs, and are recognised for direct progression to higher education and employment, as well as for immigration purposes. These qualifications describe their levels both in terms of the RQF and the CEFR. The CEFR levels are used as a proxy for describing the required language skills for progression or immigration. For example, B2 level across all of reading, writing, speaking and listening skills is typically required of international students for entry to UK universities. For a sportsperson visa, A1 level in speaking and listening is required, while a Minister of Religion visa requires B2 level across all four skills. Skilled worker or health and care worker visas require B1 level across all skills, as do applicants for permanent residency or citizenship in the UK. What this also suggests is that English language skills of speakers of other languages at B1 or B2, demonstrated via some ESOL International qualifications, are deemed to be at an acceptable standard to enable them to engage in skilled work and academic study, respectively. They are not required to provide further or alternative evidence of their knowledge of English, after being admitted to this country, by taking additional qualifications that are otherwise designed for speakers whose first language is English, such as FSE or GCSE English.

Back to qualification purposes

Using Newton's (2017) perspectives framework for thinking about assessment purposes, the fact that the SfL qualifications should help with linguistic integration of migrants, and removing language and literacy barriers to accessing employment and education, would suggest purposes that could be seen from an **engagement perspective**, helping to ensure that learning actually occurs. From this perspective,

[7] Calls for pre-entry courses reflect the research showing that certain categories of vulnerable learners that are likely to have ESOL needs, such as low-literate migrants, may require more time to acquire a language than the general migrant population, and may need literacy training as well as language training (cf. Rocca, Hamnes Carlsen & Deygers, 2020). Similarly, refugees and asylum seekers, including minors, may have suffered from trauma, abuse and danger, which are some of the issues shown in research to adversely affect language learning and require more support (Doughty and Long 2003, cited in Rocca, Hamnes Carlsen & Deygers, ibid.).

the qualification needs to exist because the amount, direction or quality of engagement that it stimulates and shapes is personally, institutionally and/or socially valuable.

Given their required alignment with the core curriculum, it appears that some of the purposes relate to the **expertise perspective**, providing evidence of a level of mastery or the acquisition of expertise of the specified content domain, as well as boosting confidence and motivation to build on skills and expand the expertise within the domain. The acquired expertise can be seen as personally, institutionally and/or socially valuable in terms of enabling individuals to assume particular roles or responsibilities in the relevant communities of practice or perhaps more broadly in the society.

What is less clear from the literature is the extent to which these qualifications serve purposes related to the **information perspective**, and which, if any, specific purposes within that. The information perspective can be seen as focusing on the use of assessment results to make decisions, where assessment results provide information that may be personally, institutionally and/or socially valuable as the basis for making decisions. As noted above, these qualifications are no longer used for high-stakes decisions that directly affect learners' legal rights. However, given apparent issues with recognition of these qualifications by employers or education institutions, it is less clear whether these qualifications provide any information that is valuable to stakeholders beyond learners themselves or qualification providers as the basis for making any decisions.

The lack of recognition of SfL qualifications by certain key stakeholders such as employers or further education institutions suggests that the SfL qualifications may not have or may not fulfil some of the information-related purposes that they should or could have. This, in turn, may be driving their own or their qualification providers' choices regarding the most appropriate language qualifications for them. Directing potential SfL learners to qualifications that might have a more well recognised and socially valuable information purpose (for instance, Functional Skills English) while not serving the engagement or expertise purpose in terms of providing expertise of the most appropriate content domain (ESOL rather than English as the first language) may be detrimental for both ESOL learner engagement and their ultimate expertise and potential for progression.

This could also be seen as a situation where a qualification justifiably has multiple (competing) purposes (cf. Newton, ibid.), but where there is not enough clarity about their relative importance to different stakeholders, and no simple answer to the question of why these qualifications are actually needed. This situation may indirectly impact on overall quality and appropriateness of provision and reduce the extent to which any of the intended purposes are truly fulfilled, potentially reducing the usefulness of these qualifications to users.

It would be important to seek clarity about the intended purposes of these qualifications and to clearly acknowledge whether there may be multiple purposes and their relative importance to stakeholders, as well as what the nature of interaction is between these multiple purposes. This would help considerations regarding the most appropriate assessment models, as well as approaches to regulation, and any trade-offs that might need to be acknowledged.

Study 1: Assessment review

Introduction

The focus of this study was to review high-level curriculum coverage, assessment models and assessment instruments from the 4 AOs in our sample in order to understand the range of approaches currently in use, and to consider whether there are any potential threats to comparability, reliability and validity in current approaches.

Method

The review was conducted as a desk-based exercise by an Ofqual senior manager with expertise in qualification and assessment design supported by an Ofqual researcher. Initially, a review of assessment approaches was carried out based on qualification handbooks and specification documents. Following this, a detailed analysis of 2 publicly available sample assessment instruments per AO and component at Entry 1 and Entry 3 was conducted. This was subsequently cross referenced with an analysis of 2 live assessment instruments per AO and component to confirm that these were not substantially different from publicly available sample materials. Where there were some differences, the characteristics of the live assessments were considered in the analysis.

Results

Summary of findings

SfL qualifications consist of 3 components (certified as either individual awards, or overarching certificate if all components are achieved at the same level):

- reading comprehension
- writing
- speaking and listening

Reading comprehension is typically assessed via objectively marked written papers. The papers tend to involve 3 to 4 tasks which consist of a stimulus text and a number of selected or short-response items which probe the comprehension of each text.

Writing is typically assessed via written papers typically involving 3 tasks. One of these tasks is always a form that needs to be filled in, while the other 2 tasks typically involve longer open written responses to designated topics (often an email or a descriptive written piece).

Speaking assessment involves performance assessment usually through 3 to 4 tasks. Usually, one of these tasks involves a candidate speaking uninterrupted about a particular topic or might involve exchanging personal information with the assessor. This is then typically followed by a role play on a given topic, which might involve the assessor and the candidate, or 2 or more candidates. At Entry 3, there is usually a task that involves some form of discussion about a given topic, often with the assessor, and sometimes with other candidates.

Listening comprehension of the interlocutor in face-to-face communication is assessed as part of the speaking assessment. Most AOs also have additional listening comprehension tasks involving listening to audio recordings and accompanying written or oral questions probing candidates' listening comprehension ability.

The assessments for these qualifications are summative, typically delivered 'on demand' or with several assessment windows during an academic year, and with opportunities for re-sits. The candidates are normally entered for assessments when deemed ready by their course provider.

The written exams (including reading comprehension and writing, and in some cases listening comprehension) are delivered in paper-based format. The speaking exam is typically delivered face-to-face in qualification providers' premises, though, more recently, it has been increasingly delivered online, via video link.

All exam papers and assessment tasks are externally set by the AOs (although some allow centres to modify some of the tasks). All AOs in our sample implement a competence-based approach (CBA) to assessment, in that they specify detailed learning outcomes (LOs) and assessment criteria (ACs) for each component/unit. This approach is consistent with the requirements of the QCF framework, which was in operation when these qualifications were accredited (for more details and discussion on CBA see Newton, 2018; Newton & Lockyer, (2022).

The ACs specified in this approach represent the backbone of assessment specifications.[8] The ACs are linked to the curriculum content statements, typically sampling the curriculum fairly exhaustively. In terms of the coverage of specific

[8] Depending on the nature of the assessment, the ACs may serve different purposes. In reading comprehension written exams, they are largely used to facilitate content sampling and item writing, rather than for describing grading criteria (Newton, 2018: 41). The latter use is more obvious in speaking and, to some extent, writing assessments.

curriculum statements, the AOs appear to be highly similar, according to our initial curriculum mapping based on their published specifications. There are a small number of differences where certain AOs choose not to specifically assess certain curriculum statements.

Beyond these similarities, which themselves are not without exceptions, our review demonstrates that there is a range of different assessment strategies and approaches implemented across the AOs in our sample, and that there are a variety of 'hybrid' approaches, with some elements inherited from the QCF era, and some perhaps influenced by the traditions of ESOL International/EFL approaches or approaches to assessment of other subjects within specific AOs.

Some of these approaches may be more or less appropriate or defensible in terms of validity, while some may be remnants of historical (regulatory) approaches and frameworks that may not entirely suit the construct of assessment. In addition, there are potential threats to comparability between SfL qualifications springing from the sheer range of approaches. These differences can be grouped into the following broader categories:

- Overall approach
    - use of internal (centre-marked) vs. external (AO-marked) assessment
- Grading
    - use of mastery vs. compensatory model of aggregation
    - requirement to pass each task vs no requirement
    - requirement to pass hurdle questions vs no requirement
    - types of ACs and/or marking criteria, for example, binary (met/not met) vs. graded (best-fit)
    - component percentage pass mark differences
- Assessment design
    - task and response strategy complexity
    - use of role play and/or discussion between candidates vs. between candidate and assessor only
    - use of role play/discussion between pairs of candidates vs. 3 or more candidates
    - use of written vs. oral questions to assess listening comprehension
    - varying number of questions in reading and listening comprehension assessments
    - predominant item types in reading comprehension tests (selected vs. open-response)

    - the extent of the guessing factor in selected response items in reading and listening comprehension assessments, both in relation to the overall pass and individual criteria
    - language demands (of stimulus materials and questions)
    - amount of language production required (in writing and speaking)
- Assessment delivery
    - whether centres can amend the tasks or not
    - whether assessment has to be completed in one sitting or not
    - time allowed per task and component
- Quality of some approaches, assessment instruments, etc.:
    - varying quality of questions in reading and listening comprehension assessments (implausible distractors, ambiguities, demand targeting)
    - clarity of mark schemes/marking criteria for writing and speaking assessments
    - extent of task administration guidance for speaking and listening assessments

With respect to curriculum coverage, there was evidence of the content areas included in the curriculum being differentially interpreted in terms of how much weight/emphasis is placed on certain aspects, how they are contextualised or how they are addressed with specific assessment approaches. A detailed curriculum alignment review is presented in Study 3 results section.

The range of approaches is perhaps unsurprising given the lack of clarity with respect to SfL qualification purposes, as well as given that there are no regulatory requirements for the AOs to implement any specific approaches, as long as these comply with Ofqual GCR. The GCR do not specify design requirements for qualifications in detail, thus permitting variation within the bounds of appropriate validity considerations for individual qualifications (Newton, ibid.).

It should also be noted that a degree of flexibility in assessment approaches may be beneficial to learners with such diverse needs and backgrounds as SfL learners and may contribute to overall validity of their assessment results even where it might pose some threats to comparability. This was also highlighted in the AO interviews (Study 2) and has been acknowledged in previous research (Lockyer & Cadwallader, 2020). Flexibility in assessment models may also benefit centres, depending on their resources and expertise (for example, centres might choose an AO offering an external assessment model because internal assessment might take up too much of their resources), although, arguably, such choices may contradict what some learners might find more suitable.

Some careful consideration of learner needs, qualification purposes and their relative weight as well as centre motivations, is necessary before suggesting 'one size fits all' approaches, even where certain threats to comparability may remain. Nevertheless, the different assessment models should be sufficiently justified in terms of validity and best assessment practice, as well as appropriate for the stated qualification purposes.

We discuss some of the more prominent discrepancies observed between AO approaches below, grouped into sections according to the broader categories outlined above.

Overall approach: Internal vs. external assessment

Some AOs implement internal (centre marked) and some external (AO marked) summative assessment models across all their components and levels, while some have different models for different levels, and sometimes for different component. The internal model involves a greater degree of administration flexibility compared to external, with assessments scheduled when candidates are ready and both candidates and assessors are available, rather than during specified assessment windows, which are more common in external assessment models.

Within the internal model, there are differences in the details of implementation between the AOs in our sample (for instance, how much flexibility in the choice of topic there is, whether assessment has to be completed in one sitting or not, whether, if completing in multiple sittings the candidates can see all parts of the assessment in the first sitting or not, whether the same task may be continued on different occasions).

The practice of internal administration of summative assessment as discrete tasks in multiple sittings has been observed in other research into internal assessment within the CBA tradition in the VTQ sector. Lockyer & Cadwallader (ibid.) cite studies finding evidence of an assessment-driven approach with BTEC (Carter & Bathmaker, 2017) and GNVQ assessors (Garland, 1998; Ecclestone, 2002; Wahlberg & Gleeson, 2003). In this approach, the assessors were focused on "getting students through" (Carter & Bathmaker, ibid.) through extensive formative feedback and the use of numerous discrete and accessible tasks. Some authors suggest this can create a scenario where candidates are coached to comply rather than learn (Garland, ibid.; Ecclestone, ibid.). Lockyer & Cadwallader did not replicate these findings in their own study, and even found that in some vocational sectors there was an emphasis on delivering learning above and beyond the course requirements to best prepare learners for employment. However, they did find evidence of a variation in the level of support and feedback provided to learners when completing internal summative assessments.

While our research did not specifically focus on this aspect nor investigate centre practices, the assessment approaches available to centres for some SfL qualifications may lead to some of the above-mentioned practices. This situation might be of more concern where the same qualification, such as SfL, can be assessed using either external or internal assessment. In such a situation it would be important to have sufficient confidence in the controls around each assessment approach in order to reduce disparities in candidate experience and potential threats to comparability of standards.

Grading

There are different approaches taken by the AOs to establishing pass grade cut scores/standards on SfL assessments. Consistent with the CBA tradition, most AOs implement a mastery measurement model, where overall competence is defined as competence across all of the specified ACs, and where competence in each AC has to be demonstrated a minimum number of times and to the required (minimum) degree. However, some AOs implement a compensatory model, without a requirement for each AC to be met.

In addition, among those that implement a mastery model, some define the pass standard purely in terms of each AC being met once across the assessment. Some define pass thresholds at individual task level, alongside a requirement for some hurdle questions to be answered correctly in the reading assessment. Some use a semi-compensatory model where, in addition to the requirement for all ACs to be met once, a numerical pass mark that is compensatory must also be achieved in some components.

There are potential validity issues with some of the above-mentioned approaches. Sometimes the models where the pass threshold is defined by each AC being met once only can result in cut scores that might be considered too low, for instance, in some reading comprehension tests (see more on this in Study 4). In other cases, the use of hurdle questions without clear construct justification may create inappropriate barriers for achievement of the pass grade. If mastery measurement models were deemed appropriate given the construct of assessment, then it would be necessary to implement them in ways that reflect best practice in using such models, ensuring valid assessment results. It is questionable whether mastery models with very low pass grade standards genuinely represent reliable evidence of mastery.

This kind of variability in measurement models has been observed in previous research into grading in vocational and technical qualifications, which are all originally based on the CBA (Newton, 2018). Arguably, this variability (assuming that individual models were implemented validly) may be more justified across qualifications with different subject matter, which constituted Newton's sample. It is less obviously justified in the context of qualifications which are based on a specific

shared curriculum and the same construct, such as SfL. Different approaches to defining the pass standard could interact in important ways with the construct of assessment, potentially unnecessarily leading to disparities in standards or disparities with respect to interpretation of results.

Another question that this situation raises is whether the compensatory or the mastery model are more appropriate for measuring language acquisition more generally. This is a high-level issue crucially related to how the construct of assessment is defined more broadly and is also potentially related to assessment purposes. As Newton (ibid.: 77) notes, CBA and its mastery model of aggregation may be an inappropriate template for those qualifications whose domains are not characterised by large sets of **critical** (emphasis added) micro-competencies, or for qualifications that would be better served by a measurement model more akin to compensation or configuration. It should be possible to determine which of these measurement models is best suited to the nature of the SfL construct and content domain as well as the purposes of these qualifications rather than perpetuating certain models for historical or other inappropriate reasons.

Assessment design

Across all components, but particularly in writing and speaking, there is evidence that task demands between assessment versions within one AO and component may not be sufficiently comparable in terms of topic demand, vocabulary complexity, and other aspects. There are also differences in reading stimulus text complexity between AOs. For instance, some texts include noticeably fewer complete sentences at Entry 1 than others, or fewer paragraphs than others at Entry 3.

There is also disparity between the AOs in the number of questions in the reading comprehension assessments, with some including noticeably fewer questions than others. This could have an effect on reliability and classification consistency (especially in interaction with the high guessing factor of some items or implausible distractors).

Duration of writing and speaking assessments also varies, sometimes considerably. Furthermore, there are differences in the guidance regarding the number of words and/or sentences that the candidates are required to write. At Entry 3, for example, the amount of writing required varies by more than 100 words between some AOs.

Speaking assessment approaches also vary in several respects. There are differences in the amount of speech that is required from candidates at the same level between AOs. In addition, some AOs require candidates to produce a monologue at each level, while some do not. There are further differences in the duration of the required monologue between AOs that do require it.

There are important differences in the use of role play and discussion in the speaking assessment too. Some AOs require candidates to only interact with the assessor/tutor across all tasks, while others require them to interact with other candidates too in some tasks. Within that, some only require interaction between 2 candidates, while others require 3 or more candidates to interact. There are advantages and disadvantages to each of these approaches in speaking assessment, but the demands that they are likely to place on candidates and the nature of performance that is elicited by each task type, as well as the demands on assessor consistency, are likely to be different.

Listening comprehension is assessed alongside speaking, and some of the listening comprehension ACs are assessed via the speaking assessment (for instance, understanding the interlocutor). However, at both levels reviewed, some AOs also include a separate audio comprehension task, while others do not. This is discussed further in Study 3 in relation to construct validity and curriculum alignment of these approaches.

Most AOs assess across different topic areas within each assessment component, meaning that a range of topics is sampled in each component and across components within one level (for instance, family-related issues, shopping, course application, news articles on different topics, advertisements, tourist guides). In contrast, some AOs pre-define a range of topics and each component is assessed within each of those topic areas only rather than across topic areas. For instance, a whole reading comprehension paper would relate to the context of a single topic. This is intended to allow centres to choose the appropriate and relevant topics for the candidates to be assessed on, though with the requirement that at different levels a candidate is assessed on different topics. While there might be good reasons for this given how diverse the SfL cohort of learners is, it might limit the curriculum in terms of the range of vocabulary covered and evidenced at each level.

Assessment delivery

Several differences in assessment delivery were noted, mostly with the AOs that implement the internal model of assessment. Those AOs give some flexibility to their centres to amend the reading comprehension tests in particular, but also offer more flexibility in relation to the topics of speaking and writing assessments, compared to the AOs implementing the external model. In relation to potential amendments of the reading tests, it was noted in our interviews (see Study 2) that this opportunity is rarely taken by the centres, however, as the work involved to implement changes is usually prohibitive.

Among other flexibilities, the option to complete an assessment in multiple sittings is offered by some AOs but not others. Potential issues with such flexibilities were discussed earlier in the context of internal assessment approaches more generally.

Some variation was observed in relation to the time allowed by different AOs for different tasks or components, which could have some implications for task demand and comparability of candidate experience. For example, for Entry 3 reading tests, AOs varied between 2 and 4 minutes per mark. In Entry 1 writing assessments, the time allowance across the 4 AOs varied between 10 minutes and 20 minutes per task, with some allowing around 15 minutes per task. There were similar discrepancies in the speaking component.

Quality issues

A range of quality issues were noted with the questions in reading comprehension assessments, and, to some extent in the listening comprehension assessments where these were designed as written tests. It should be noted that our reviews did not involve statistical analyses of item and test functioning as we did not have access to AO's data for this research[9]. Our qualitative review, however, strongly suggests a need to obtain evidence from the AOs that their written tests function appropriately.

One of the most common issues noted with some tests was the potential for getting the selected response items right by guessing. The guessing factor was variously increased by a high proportion of true/false items in some tests, or implausibility or cues in distractors of multiple-choice questions in others. On the other hand, some tests were designed effectively to reduce the potential for guessing. Instances of items with implausible distractors, ambiguities, and apparently too low or too high demand for the level of the qualification were noted across most reading and listening comprehension tests to varying extent.

Our review also found that there was a range of approaches to and level of detail in assessor guidance in relation to the administration of speaking assessments. There were also varying approaches in relation to how assessment criteria were elaborated for both writing and speaking assessments. Some external assessments appear to provide more guidance than some internal assessment. This is of concern as internal assessors are likely to depend more on this guidance compared to external assessors. This finding echoes previous research which identified potential vulnerabilities in the delivery of internal assessments related to assessor guidance and training, and their potentially differing interpretation of AO-set standards (Lockyer & Cadwallader, ibid.).

Issues with a lack of detail and clarity of mark schemes and criteria were also noted with some external assessments. While it is possible that further clarity is provided to

[9] See, for instance, Black, He & Holmes (2017) and Black, He & Morin (2019) for examples of statistical evaluation of some vocational tests.

assessors during standardisation and training, arguably it would be preferable to provide clearer, publicly available criteria as far as possible. Furthermore, variability in assessor standardisation itself may pose threats to validity and comparability, particularly in those approaches that involve the mastery model for specifying grade standards (see Newton, 2018.: 69).

Study 2: AO interviews

Introduction

This study sought to explore the perceptions and views of 6 AOs who offer SfL qualifications. The overarching research questions for these interviews, outlined below, are deliberately broad, allowing us to explore the perceived purposes of the qualifications and how they are developed and delivered to meet those purposes.

1. What do AOs perceive as the purposes of SfL qualifications and who, therefore, are the learners that AOs seek to support?
2. How do AOs who offer SfL qualifications develop their specifications and assessment materials with regard to their purposes, the core curriculum and any other external frameworks or guidance?

Ofqual regulates such that qualifications are valid with respect to their established or known purposes and reliably indicate the knowledge, skills and understanding students have demonstrated. Understanding the purpose(s) of a qualification provides the foundation from which to understand the validity of the assessment (and, as part of that, its reliability). The above research questions therefore complement the wider research report by providing important context for understanding ESOL SfL qualifications.

Method

Given the open and explorative brief, a qualitative approach using semi-structured interviews was selected. Group interviews were conducted with small groups of employees from 6 different case study AOs.

Interviews typically involved 2 or 3 Ofqual colleagues and between 2 and 5 employees of the AO. The AO selected their own panel of interviewees, seeking to provide expertise across all areas of the development and delivery of their SfL qualifications. Interviews were conducted using a flexible interview schedule and took place via video conferencing software, typically lasting for 2 hours. Each interview was recorded (video and audio) and the audio elements were transcribed

verbatim for analysis. The interview schedule was developed to explore the AOs' perceptions of the purpose and constraints for SfL qualifications within the market, their broad processes for the development and delivery of qualifications within that context, and their broad processes for setting and maintaining standards.

A flexible thematic approach was taken to analysing the data, with the goal of identifying patterns across the 6 interviews and making comparisons between them in relation to the research questions (Braun & Clarke, 2006). We did not therefore conduct a full 'line by line' thematic analysis of the data and instead undertook a broader comparative analysis of our case studies. The lead researcher used NVivo qualitative analysis software to code the written transcript with patterns and themes. The researcher had access to the audio files while doing this, allowing them to clarify the tone of an extract, or otherwise verify the accuracy of the transcription, where necessary.

Some of the codes that were used within the process had already been broadly identified following discussion between the Ofqual staff who attended the meetings, while others emerged more organically through the analysis process. The researcher usually coded sections of the transcripts rather than individual sentences, the aim of which was to take account of the context of a comment and to avoid analysing statements in isolation. The researcher revisited transcripts they had initially coded to ensure that the coding scheme had been applied thoroughly and consistently, focusing on transcripts that had been coded earlier in the analysis process at a point where the coding scheme may have been partially incomplete. At the end of this process, a second researcher scrutinised a sample of the transcripts to confirm that they agreed with the coding and provide challenge where they did not.

It is worth noting that this research does not capture the views of all AOs who offer ESOL SfL qualifications but rather it seeks to illuminate the key issues and how they may be influenced by the context of the AO. This will help us to gain a more nuanced understanding of the perceived purpose for these qualifications and how that influences how they are developed and delivered in practice.

Key themes emerging from interviews

Below, we outline the key themes that emerged from the analysis. We have selected quotations to illustrate these themes. We begin with a short summary of the overall findings.

Summary of findings

It is clear from the views expressed in our interviews with AO staff that the SfL qualifications are seen as having enormous value to learners and society as a whole.

The AOs emphasised the breadth of purposes of SfL qualifications as well as a wide variety of learner background and needs.

There was a sense, however, from all AOs that these qualifications were undervalued in terms of funding. They also said the qualifications were not sufficiently recognised for their potential to enable learner progression, for instance to apprenticeships, or T Levels, particularly in comparison to Functional Skills English (FSE) qualifications. Practitioners told us that this affected stakeholder attitudes towards the qualifications, including centres, employers, higher education institutions, and learners themselves. They thought that this was also partly to do with a lack of understanding, on the part of some stakeholders, that ESOL qualifications have an enabling function in terms of developing broad language skills of learners with ESOL needs, from very basic to advanced, rather than just focusing on literacy skills.

They emphasised the need to recognise the unique value of these qualifications for addressing ESOL learner needs, while recognising their parity with similar qualifications such as FSE. The SfL qualifications were not seen as dissimilar in their enabling function to ESOL International qualifications, some of which are widely recognised at competence levels similar to Entry 3, Level 1 or Level 2 by reputable higher education institutions as well as for job recruitment.

The AOs discussed the flexibility of the SfL qualification structure, its modularity and the variety of assessment models that can be used as beneficial and justified by the needs of cohort and centres. They described how learners and centres required flexibility of choice and adaptability of assessments to accommodate the wide range of contexts in which these qualifications are delivered. All our interviewees referred to the flexibilities offered in the context of speaking assessments in particular. However, AO staff in our interviews suggested that the flexibilities offered to centres in terms of the possibility for written reading comprehension assessment adaptations are rarely taken up, although the AOs themselves offered some choice of context across different assessment versions in some models. Furthermore, the choices that centres make, for instance, in relation to using internal versus external assessment models, do not seem to always be tailored to the learners, and sometimes appear to be driven by centre needs (for instance, availability of assessors).

Our interviews with the AOs suggested that some of the assessment design choices, noted in our assessment and curriculum alignment reviews as potentially sub-optimal or not entirely appropriate for the construct of assessment, were partly historical, and linked to the requirements of the QCF (the previous regulatory framework). These design choices were not always considered by our interviewees to be reflective of best practice in language assessment (for instance, atomistic assessment criteria, certain content aspects enshrined in the units, choice and implementation of measurement models, weighting of certain curriculum aspects). However, those we

interviewed thought that the QCF has provided a helpful mechanism for ensuring a degree of consistency of approach, construct and standards across AOs, particularly because it allowed the units of qualifications to be interchangeable across AOs. This was seen as important for learners who often need to move and complete different units with different centres and AOs, though there was acknowledgment that other mechanisms may exist for achieving this.

The AOs were keen to emphasise that, although they saw the SfL qualifications, as well as the core curriculum on which they are based, as generally effective and appropriate, there was room for improvement and suggested aspects that may benefit from it. Within that, though, they strongly recommended retaining sufficient flexibility in qualification design and assessment models, which would ensure that there were no additional barriers to participation and learner progression.

Breadth of purpose

The AOs we interviewed were keen to explain that there is no such thing as a 'typical' ESOL SfL learner. Learners appear to choose SfL qualifications for a wide range of reasons.

> It's really for everyone, there's no such thing as a typical ESOL learner. It's people who want to stay in the country and to progress and to integrate really into British society. So, some of them may have been here for a while and have suddenly decided, yes, the children are going to school now, and I need to be able to help my children, or it might be that they now want to get into work, or […] they've been told that they will progress more if they can speak better English, it's really all sorts of things. AO4
>
> Learners will take ESOL qualifications for integration purposes, for job or career development, for further development in education, for helping their children in school provision, for their day-to-day life, for social reasons. There isn't one specific reason why people would take them and quite often it's across the range and it's predominantly obviously to enhance their life. AO1
>
> So, whilst it is predominantly a language qualification, the ESOL teaching community is very much versed in the wider issues and the wider needs of people learning English as a second language. So predominantly what these people want, people taking ESOL qualifications, they want to be able to take a full part in working life. They want to be able to live independently in the UK without the need for interpreters or translators, whether that's informal family or formal services. […] But we also have secondary schools, sixth form colleges and further education where we have learners between 14 to 19 or 16 to 19 on study programmes, and they're learning English in order to help them achieve in other qualifications, such as GCSEs, functional skills, which may be a progression

> route for them, A levels, routes into higher education, apprenticeships, a whole range of academic opportunities. AO3

The AOs also noted that ESOL learners are a highly diverse group in terms of their background, and that the cohort profile can fluctuate over time, partly depending on geopolitical factors.

> As an example, settled minorities, economic migrants, whether from Europe or whether from the rest of the world, we have a lot of refugees, asylum seekers. We have foreign victims of trafficking who've been brought into the country, or modern slavery, brought in against their will, and a lot of younger learners too. Most of these are adults, most of the ESOL cohort is 19 plus. AO3
>
> I would say that we've seen the types of learners flex according to what's happening in the wider world, so in recent years we've seen a larger influx of refugees who needed a lower level and in fact ESOL starting at entry 1 wasn't always accessible to them, they needed something below that level, so we've seen quite a big increase of that recently and yet that has changed, we've seen a bit of a circle, and it depends on where they are in the country as well because of local areas tending to focus their support on different things. AO4
>
> […] I think maybe more younger learners, if we're talking over 20 years, there's more access by 16 to 19s. AO1

The AOs noted that ESOL learners vary considerably in terms of their education and employment background. For example, while some learners may not be literate in their native language, others may be highly skilled professionals who wish to develop their English to improve their job performance and further their career. Learners also vary in terms of how much direct interaction they might have on a daily basis with native English speakers, with some not getting much interaction beyond their ESOL classes.

> I have had in the same class people that you are literally teaching how to hold a pencil because the only time they went to school was under a tree somewhere in their village, sitting next to somebody that might have a PhD and just because they have lack of English doesn't mean they have lack of intelligence. AO1
>
> […] they may be well-educated Europeans who simply want to gain, for example, a medical qualification in the UK and therefore need to develop the language and skills in order to be able to do that. So economic migrants such as doctors, nurses, anyone in the healthcare profession […] and in all sorts of different professions. AO3

> […] a lot of the time we're talking about situations where ESOL learners haven't had the opportunity to speak English in an environment, in many cases they live in an area where it's very easy to fall back into your mother tongue. AO5

Our interviewees told us that ESOL SfL learners completed their courses in a wide variety of settings, from sixth form and FE colleges through to prisons and community groups. The latter of these, community groups, appeared to be particularly important for some AOs. The emphasis, from a teaching and learning perspective, is on providing a reassuring and supportive environment for entire community groups who were seeking to learn English, with ESOL provision being integrated within wider educational and social programmes.

> I think the community provision seems to be the key for [us] and where our audience is, in that it's delivered in a community environment where the learners tend to be of a similar ethnicity within a class size and generally a similar age. AO6

> …a lot of it is delivered within community groups where people feel more secure, where maybe the teachers share a first language with the learners, where there is a lot of support, not just language development, and often that helps to get across the barriers. […] So, we have things like, in the voluntary sector there are things like cookery clubs where trained professional teachers will actually integrate the cookery with the language teaching and that way people feel safer and more secure to participate and achieve. AO3

Given the breadth of purpose for ESOL SfL qualifications, and the diversity of learners taking them, the AOs told us that they tended to form close and dynamic relationships with their centres to support them in meeting the specific, and often changing, needs of their learners.

> The Adult ESOL Curriculum and the other frameworks that govern these qualifications kind of lay out what our rules are. So, the navigation and negotiation happen in the classroom and that's where we often have to support the teachers with respect to how to help their learners get to a point where they can travel through the gateway that we have set up. AO2

> So there are centres that have been doing things that focus on childcare because they know that's what's relevant and we do have centres more and more who say to us, can we do ESOL in a vocational context […] So we were having discussions with a very large college about how we could support them in terms of adaptation, not needing new quals because they still have to use the same language structures, it's still the same ESOL Core Curriculum, but putting in contexts that make them more relevant AO5

> […] we do a lot of networking activities and a lot of centre-facing support. We have a plan of centre-facing support activities, so, much of our knowledge of our centres and their cohorts comes from there, as well as the changes, because learners come and go and patterns of the types of learners really do change, and I would just say it has flexed over time. We've seen all sorts of cohorts come through and need ESOL for slightly different reasons and it has been able to adapt to that, which I think is probably quite important and, similarly, I think a lot of our learners or candidates use the qualifications very flexibly. AO4

This flexibility was considered particularly important and is discussed in further detail under a separate theme below. Overall, it was clear that those we interviewed felt that ESOL qualifications had a broad remit and could therefore meet the needs of a wide variety of learners.

Language as a key facilitator

The AOs were keen to stress the important role that language skills play in providing people with opportunity. Language skills were considered to be a bridge into society, allowing individuals and groups to integrate better with their communities and to access education and employment, helping them to fulfil their potential. Indeed, ESOL SfL qualifications were considered vital for unlocking the potential in learners commensurate with their capability in their native language, allowing them to use existing skills for employment in the UK where previously their lack of English language skills had been a barrier.

> They may well start off in an ESOL class thinking I need to learn just to be able to interact and things, and then suddenly think, actually, this can take me a lot further. And they'll go onto a different educational course and that may lead them into education, and they'll end up doing something that they never envisaged, just because it empowers them to move forward with their life. AO1

> […] many of those learners come to college with quite low levels of English language but quite high ambitions about what they want to do, whether it's to go to university or do an apprenticeship or something similar. And they run a study programme where they will actually take their learners through the ESOL courses in order to give them the language skills to be able to access GCSEs in their second year, you know, in a subsequent year of the study programme, or the access to HE qualifications that will enable them to then move up. And these are young teenagers or 14- to 19-year-olds who come from a range of different backgrounds and for whom without that language development as part of their study programme they would not be able to achieve the other qualifications and they would not have the progression opportunities. AO3

> So, if you're, for instance, an engineer from the Sudan, it would be, in my view, in the government's interests to get to the point where we could access your engineering skills, not just your ability to be a taxi driver, personally. AO2

All RQF levels of the ESOL SfL qualification were considered to have a purpose for language learning. The Entry level qualifications were considered to be an important foundation for some learners, helping them to master the basics and gather confidence for further study. They were deemed justifiably divided into 3 sub-levels. On the other hand, access to specialist ESOL qualifications at levels 1 and 2 was also seen as important for unlocking higher-level skills as well as, consequently, employment and other progression opportunities.

> Actually, when you're learning a language and you know nothing about it, not even a single word, it is a continuum, but it is a necessary continuum. Even if you have a really high goal and some high aspirations, you're really aware that you cannot achieve and attain those goals and you need to take everything a step at a time. So certainly Entry 1 qualification probably cannot unlock great job opportunities, but maybe it can unlock some when there's no interaction with someone, with a native speaker or with people speaking the target language. I would expect that Level 1 and 2 and Entry 3 qualifications would unlock different employment opportunities, but I would not be able to categorise in terms of importance the Level 1 and Level 2 qualifications and say that the entry qualifications are not as important. They are an important steppingstone. AO2
>
> […] I think this is absolutely a qualification area that is needing that recognition of the 3 [Entry level] sub-levels. AO5
>
> I would argue that ESOL Level 1 and ESOL Level 2 better equip learners for working life, because it gives them the grammatical structures, the formal versus informal language, the vocabulary, the range of comprehension and understanding that they need to perform in technical training, higher education, skilled employment progression opportunities, whereas if they're only at Entry 3 then their education and training and work opportunities are limited to lower skills. AO3

As hinted in the quote above, AOs highlighted the need for ESOL provision at higher levels and emphasised that learners at higher levels do not benefit in the same way from FSE courses.

> So sometimes we're seeing people who should be doing ESOL qualifications probably doing other qualifications, is what I mean to say. So, we often see people moving towards Functional Skills where probably application is more the purpose than language acquisition. So that causes some challenges I think for learners. AO4

> […] in my experience not only do ESOL learners not fit functional skills, it's square peg, round hole, […] it just doesn't work, a lot of ESOL learners at Level 1 and Level 2 are actually at a higher level with their functional skills and quite often their reading and writing skills than native speakers doing Functional Skills. AO1

There were suggestions across all AOs that the current funding landscape, particularly for higher levels, may be driving some of the patterns of provision, where ESOL learners may be directed onto FSE courses for funding reasons. This was seen as limiting the potential for ESOL SfL qualifications to be as valuable and impactful as other qualifications at similar levels (for instance, FSE).

> In my opinion there should be parity between functional skills and ESOL Level 1 and Level 2, because second language learners are disadvantaged by the lack of grammar and vocabulary content in the functional skills qualifications as they currently exist. Now, this also is a funding issue really, you know, because functional skills is part of the entitlement but ESOL is not and therefore employers don't necessarily recognise ESOL Level 1 and Level 2 as having parity with functional skills because it's not part of the entitlement. AO3

> […] providers didn't necessarily choose the qualification that was right for the learners, they chose the qualification that was more accessible for them in terms of drawing down funding and things. It's a lot easier to deliver a qualification where it's all being funded by the government than chasing learners for money. AO6

There were clearly some concerns about the nature of the qualification landscape, specifically around funding, and how this might affect take up and progression. There were also concerns around ensuring fitness for purpose of the qualifications that were made available for learners.

The core curriculum and the NSAL

The AOs were generally positive about the nature of the core curriculum and suggested that it was a helpful tool to inform assessment context and standards. The curriculum was also viewed as an important resource for teachers and learners for guiding and gauging their progress, as well as for teachers' skill development.

> The core curriculum is the bible for anybody really involved in ESOL right from, I mean, when you do your teacher training you learn about the core curriculum and then when you go on to teach you use the core curriculum, knowing that they're going to be assessed against the core curriculum. So, it's just the foundations for everything and you know what sort of level the learners are going to be at, whether you're teaching or assessing. AO1

> The simplistic answer is that those frameworks, the core curriculum and the adult literacy frameworks, informed the specifications, the test specifications. And then we are governed by our test specifications, which are our interpretation of those other documents. AO2
>
> […] learners do appreciate the fact that there is a curriculum. They do really like the fact that there is a framework and there is a clear progression route and, irrespective of which awarding body or which college they're in or which provider they're in, they can see their progress in a very real and tangible way. And I think it's absolutely essential to have that as part of a national strategy for ESOL. AO3

The benefits of the curriculum were perceived to extend to helping ensure a consistent approach was taken across qualifications and over time.

> You've seen our specifications […], there's an entire table at each level which says, this is the criteria, and this is obviously from the curriculum it comes from, so we have been as faithful as we can be, partly to support the teaching and learning process, but also partly because we want to be able to ensure that walk across from our qualifications to any of the other AOs and, as I said, this sharing of terminology and a common language, that's the way you ensure it. AO2
>
> […] when we wrote the initial samples, they were all very clearly based on the core curriculum, the learning outcomes mapped to the core curriculum and every paper links back to it. So, when our writers are writing the assessments, they know that they've got to follow the same template, we've got extensive guidance which reminds them about the core curriculum, which reminds them about our requirements, which reminds them about contexts so that they are comparable… AO5
>
> Yeah. Nothing's perfect, but on the other hand it's an entity that's there and I think that's why in many ways these suites of qualifications are successful and why there's a group of them. So, a learner who goes from one college and says they've got an entry 3 from one AO and then they go to another college that does another AO, they know what that means. There is this strong comparability […] it's like the bedrock, but it does need to be looked at again, it needs an examination, but it's there, it's not perfect, but it's there and it's clear, people know what it means. AO1

However, views on precisely how effective this was varied somewhat between AOs. Some comments suggested that differing interpretations of the curriculum were possible, something which has the potential to lead to a degree of inconsistency between AOs.

> So, the learning objectives and the assessment criteria are directly based on the core curriculum. There was a lot of discussion about what they meant in relation

> to the core curriculum and the relationship, there was a lot of teasing out of what the writers of the core curriculum had intended. AO1
>
> There of course are arguments for having an explicit assessment of listening and I would say that we decided for ESOL Skills for Life to prioritise two things: one is the interactive listening with someone who is in a slightly different social position from you, that relationship thing, and the other is the discussion among peers where (a) you have a decision-making responsibility, (b) those social layers are removed, but you have the overlay of having to manage perhaps accents that you're not as familiar with, varieties of English you're not as familiar with. So, we chose to prioritise those within the time constraints of the exam. […] there are also aspects of that curriculum that justify our decision […] So we felt that this is better or more easily operationalised through a live context instead of having the typical recording and listening to the contrived recording twice or three times or as many times as you feel. AO2

Those interviewed felt the curriculum was fundamentally fit for purpose, if in need of modernisation and refinement. The general view was that evolution over revolution would be recommended. The AOs thought that there was certainly room to improve the curriculum but that this could be achieved with careful revision rather than needing a fundamental reform.

> Well, the Adult ESOL Core Curriculum in itself can be either a barrier or a support. So, when there is such a close relationship between an assessment and the curriculum on which it's based, that symbiosis is affected by whoever is in charge of making decisions. So, when [it] is next reviewed and revised, it might actually be a really good idea for the assessment organisations to be included so that it is learning, teaching and assessment, rather than just learning and teaching. Because then we can actually fold in some of the really new approaches to thinking about the role of assessment and bring in concepts like assessment for learning, which I think are really exciting because they have the potential to engage the learner much more in their process. AO2
>
> Yeah, I think some of the referencing to the digital world is a tiny bit out of date. That crops up quite a lot. I think it would be really nice to make more reference to employability in it, that would be really helpful, and actually just broadening out, in the examples broadening out some of the contexts would be really helpful. AO4
>
> So, I think if you were to say to me, should the ESOL core curriculum change, I would say probably it's more a tweak than anything else. […] maybe it's some of the actual examples that need changing and activities, […] it needs a refresh. AO5

Some AOs pointed out that the link between the core curriculum and the NSAL is a complex one. On the one hand, the nature of the ESOL cohort, with many learners

not having basic literacy skills in either English or their first language justifies explicit emphasis on teaching and assessing literacy within ESOL SfL qualifications. On the other hand, the AOs were keen to point out that ESOL SfL qualifications go beyond teaching and assessing literacy, because the language learning needs of the ESOL cohort are much broader and very different from those for whom English is their first language. Some AOs drew our attention to the link between the core curriculum and the NSAL with regard to issues around the recognition and parity of esteem of ESOL SfL qualifications and other English qualifications.

> I think the Adult ESOL Core Curriculum followed on from the national literacy standards and that move was to, it was actually to increase the standing of ESOL by putting some parity there and not having it as ESOL qualifications or ESOL learners as a lesser thing than these national literacy standards, which were the important thing at the time. AO3
>
> ESOL learners are not being taught literacy. It may well be they're not literate in their own language, but primarily it's about language and it's about language in realistic contexts that are relevant to them. AO5
>
> I think I would still say that the core curriculum, while being linked to the national literacy standards, is not just literacy, the core curriculum at the moment is language and literacy […] we would still need something like it, not necessarily to the literacy standards, but something like the core curriculum. […] Language development is very important, and it differs from other English qualifications such as GCSE and Functional Skills because there is that focus on developing a lexis, developing comprehension and developing the grammatical skills that underpin the use of language… AO3

Some AOs suggested that basic literacy, in terms of learning the Latin alphabet and letter formation, are skills that may benefit from being taught prior to Entry 1 level courses, with Entry 1 courses deemed too high level to easily accommodate learners who have no basic literacy skills whatsoever. On the other hand, it was pointed out that learning of these skills is sometimes complex and non-linear, and hence may need reinforcement and repeated assessment, particularly at lower entry levels.

> So, I think it's just a really complicated area and we used to have the pre-entry curriculum, didn't we, years ago which took more of a stepping stone, milestone-type approach. […] I think the other thing at this end, which is often overlooked, is the need to relearn. Learners will come and they will learn the alphabet, let's say, or they'll learn parts of it, or they'll learn to form letters or whatever it is, and it might be even speaking, listening or reading, but they won't necessarily remember it and they might have to redo it and redo it again even before they're comfortable moving forward. AO4

> It's important that [the alphabet is] explicitly tested at lower levels, specifically at Entry 1, because obviously it is important to differentiate between the lower- and upper-case letters. I think it's implied that learners will have that skill as they move up the levels, although it's not necessarily always the case. AO1

Several AOs pointed out what they saw as an unfortunate interplay between the use of an atomistic QCF-based approach to specifying assessment criteria and a mastery model of assessment with the approaches to assessment of literacy in SfL qualifications. Generally, the AOs were of a view that knowledge of alphabet, spelling, punctuation or text purpose may need to be assessed more holistically and communicatively than they currently are.

> It's a more holistic thing than saying I can understand alphabetical order, I can understand upper and lower case, I can understand the purpose of a text, it's more holistic than that, but that was the way the QCF we felt led us… AO3

> We know that as part of the Adult ESOL Core Curriculum there is some attention paid to spelling conventions and punctuation. We also know that in real life spelling almost never interferes with the understanding of the text. And I'd love anyone on this call to be able to say they absolutely 100% know how to use punctuation. AO2

The interplay with the QCF is discussed in more detail below.

The QCF legacy

The SfL qualifications were accredited by Ofqual in 2014 against the rules of the QCF. Awarding organisations are no longer required to adhere to the (now defunct) QCF, but it remained prominent for many of those we interviewed. As is implied in some of the quotations in previous sections, our interviewees told us that, in some cases, the structure of current ESOL SfL qualifications is a legacy of the QCF, though also pointed out that the similarities between qualifications are also there due to them being based on the core curriculum and the NSAL.

> The hangover of the QCF is still very much within our ESOL products and I think there could be a lot more efficiencies… AO4

> […] when the RQF was introduced in 2018, some of the AOs have changed their units, but fundamentally they're the same, because the content is still very coherent and there's a lot of similarities across them, because they're all based on the Adult ESOL Core Curriculum and the National Standards for Adult Literacy. […] I think when we were talking about the QCF model, that is something that we could look at. It would be a significant change. If you're going to do it, you have to do it properly. At the moment centres are happy. AO1

Some AOs felt that timescales in which to revise qualifications, following reform or a change in requirements (such as the removal of the QCF), were insufficient.

> I think that our reliance to the QCF is also lack of sufficient time to respond appropriately to the withdrawal of the QCF. So, the easiest solution was to carry some aspects of the QCF that guaranteed that comparability, for example. […] May I also add that for me another barrier, not to good practice per se, because there's not one right way how to assess something, there are different ways, but the timeframe, the way that some changes are announced and the timeframes that AOs have to react to those changes. And the fact that I used instinctively the word react, that's probably not a good indication as well, we shouldn't be reacting, we should be part of it, and we should have the time to plan it carefully. AO2

> I would just add to that, the timeframes that we get often to redevelop, design, get things out there literally up and running and working are so short and seem to get shorter and shorter. AO4

The reliance of the QCF has had a mixed impact. On the one hand, the AOs thought that the QCF provided a helpful mechanism for ensuring a degree of comparability across AOs, particularly because it allowed the units of qualifications to be interchangeable across AOs. This was seen as important for learners who often need to move and complete different units with different centres. AOs did, however, acknowledge that other mechanisms may exist for achieving this.

> Yeah, so given that this is an entirely UK-based qualification, to remove a framework within which we all work, would inevitably create the need for a NARIC-type equivalent and I think that it would be counterproductive, personally. […] And I think the AOs' reliance on the QCF, although we're not allowed to, I think that that speaks for itself for the need of something being in there to ensure the comparability. AO2

> Learners often have to move. If they're being settled or housed by a local authority, they may be moved at short notice to somewhere else in the country and so we have to have a common framework within which to teach them so that they can take a qualification elsewhere and complete it somewhere differently. AO3

On the other hand, the QCF has proven somewhat restrictive at times, with certain content aspects enshrined in the units, the atomistic assessment criteria, and a strong emphasis on mastery, which most AOs recognised was not necessarily reflective of best practice in assessing language acquisition. QCF rules also partly determine the AO approaches to setting standards for individual assessments.

> […] the framework, it was very specific about, these are your assessment criteria, and you must evidence every single one, and in our design of our assessments, if we were to design it now, I don't necessarily think that those assessments would look like that. AO4
>
> I think it was that, certainly when we first met and we met as all the AOs, there was an awful lot of discussion about how can you make a language fit into these very separate, small assessment criteria tasks; whereas a language is somehow more than that. AO3
>
> […] if you asked me what we've got that we've currently got that I wish we didn't have, I'd probably say I wish we didn't have in every single level a form-filling exercise, which is something we can do nothing about because it's in the units, it's a requirement. […] I think there are some things we would do differently – sometimes I feel it's a bit of a straitjacket because of the QCF units and maybe without that it could give us more flexibility. AO5
>
> The setting of the pass mark […] I think it was a continuation of the QCF where it required two-thirds of the content you were examined to be correct so that you could pass. So, this is the continuation and the legacy, let's say, of the QCF. […] May I also add that there's supporting skills and how they're combined into genuine communicative performance, and there's a tendency to concentrate at the level of the supporting skills. So, to try and define out at such a level of detail that that combinatory process is then forgotten, and it doesn't help anyone. AO2

Flexibility in design and delivery

Although all AOs build their assessment on the foundation of the core curriculum and the QCF, they considered the current flexibility that exists in the qualification structure and assessment models as essential to making the qualifications accessible to the full range of learners and allowing them to be as educationally valuable as possible. These flexibilities were often seen as tied in with the necessary flexibilities that were also required in course delivery.

> I think that flexibility is probably the last thing that should go in terms of regulation. I think that's really important to our learners and really important to our centres who are often small, and community based. AO3
>
> So, I think the cohorts are representative of people who do come in and out of learning, might be full-time, might be part-time, might even have to change that mode of attendance during the programme according to personal circumstances, so that flexibility has sort of enabled, the qualifications have stood up to that I would say, which I think is quite important. AO4

In particular, the AOs noted the importance of flexibility in terms of modular qualification structure, with individual units of assessment, and the ability to take individual units at different levels or move between different AOs.

> So, we offer all 5 levels, and we offer reading as a separate unit, writing as a separate unit, speaking and listening as a separate unit, or all 3 combined as a certificate. […] So that modularity and that flexibility is very much with the needs of the learners in mind. AO3
>
> I think there are certain cohorts where [lack of modularity] would create a really big barrier, so prisons would be a really good example to that, just by the very nature that people move around, and I think that would be the biggest challenge. A lot of ESOL learners do move. […] They can have a spiky profile, so that they can be higher at one skill, one mode than the other and I think that helps them on the way through. AO4
>
> You just go for it, you can get a level 1 speaking and listening and do an entry 1 reading and writing, that's your level, just that's who you are. AO1
>
> What we find is a lot of centres offer ESOL over 2 years rather than as a one-year programme. You've got to look at the speaking and listening is 120 hours, which is more than a GCSE that's delivered over 2 years, the writing is 90 and I think the reading is 60. They are, more often than not, not delivered as full-time qualifications, especially with adults. They're a daytime part-time class or an evening class kind of thing and they just come in potentially for a couple of hours a week maybe once or twice. So, it does build up and it does take time for these learners. So, we often have learners that will start off on the speaking and listening in year one and then progress on to the reading and writing in year 2. AO6

The option to use internal assessment models with their strong formative approach was also seen as important for centres and learners, especially at entry levels, by some AOs. This model provides familiarity with the tutors and assessment surroundings, as well as flexible delivery (on-demand assessment with assessors available at short notice; options to deliver different parts of a question paper on different days; flexibilities around re-sits (of whole or parts of examination papers; scope for centres to amend or contextualise assessments as appropriate, in line with AO rules; etc.).

> I think by having internal assessment with external quality assurance, it makes it easier to actually, if necessary, contextualise to different contexts and I think that's the point, contextualisation, which I think is really important for ESOL learners AO5

> I was just going to flag up, with any increased regulation I think, not just the flexibility of the delivery but the absolute essential nature of continuing with the formative approach. Particularly for recognising the nature of this candidature that the backgrounds that they may have come from, the unstable lifestyles possibly, challenging conditions, challenging lifestyles, lack of experience of education, so I think that's probably possibly a difficult one to balance, but I think achieving that balance is essential. […] Presumably it is easier to regulate if you have an end-point assessment which people do and a pass mark and that minimum thresholds have been agreed and there's a final assessment and that's the way it's done, but I think there would be a loss then for ESOL learners in having that, especially those learners maybe at entry level who are not used to formal education, are not used to assessments, who find being assessed by your teacher is a much more comfortable experience, being assessed within the classroom is a much more comfortable experience. AO3

> I think this is an important point when you're talking about E1, E2 […] that there was sufficient flexibility in the design that learners could actually do it and achieve without getting overwhelmed, so we enable them at E1 and E2 to take different parts of the paper, as long as the paper was kept confidential, at different times. AO5

The AOs that deliver external assessment also noted that there was scope for flexibility within their own model, particularly within the speaking assessment, where assessors are trained to pick up on what the relevant topics might be for a specific learner. They also noted that they provide formative feedback to learners based on their external assessments. It was also acknowledged, however, that in some respects the external model which involves fixed exam dates was inevitably less flexible than internal assessment and may not work for some learners.

Some AOs saw the flexibility and choice in terms of high-level assessment models (for instance, in terms of internal vs. external assessment) as beneficial to both centres and learners, though it was not always clear how far centre choices were always tailored to their learner needs, rather than their own resources, expertise or other motivations. It was suggested that some learners prefer the formality of an external model, and perceive it to 'hold more weight', while other learners benefit from the flexibility of timing and familiarity of setting in an internal model, as mentioned above. How far learners themselves are able to choose centres depending on which AO and model they use to match their needs remained somewhat unclear.

> I think it works quite well where you basically choose your awarding organisation on what you as a provider want and I think the learners in the end get a very similar experience. […] Part of that as well is due to the nature of the offer and how ESOL is not like the other English qualifications in the sense there is quite a

> difference of choice available for a centre, a provider to choose in terms of their assessment methodology. So, you may find that some centres want to go down an on-screen route for their reading exams at level 1, level 2 so they'll go to City and Guilds, some of them want to have a more formal speaking and listening exam so they will go to Trinity where Trinity will send an interlocutor along to do things and an assessor as well, so it's how that is. For some learners that's quite important, because obviously that's how a lot of the EFL qualifications are run and they're not run with your own teacher, so they're trying to mirror what happens there as well. AO6
>
> The other bit of course also relates to decisions that colleges make around their own resourcing and their own feelings of expertise. We all know that the tutors at these FE colleges have busy lives and if the college chooses an internal assessment model the workload for them is actually increased and also the expectations of expertise increase. So, I think the colleges are always balancing that as well. AO2
>
> Some centres definitely want external assessment. They like the formality of it, […] the event. […] Some learners get dressed up for it. They'll put on their best clothes to come in for an assessment and they like the fact that they're meeting and showing off their skills to somebody externally and that's a big part of it. […] and they feel that it holds more weight. AO1

Context and topic flexibility at task level were noted as a particularly valuable aspect of flexibility given the wide range of learner background, ages and interests, as well as the breadth of qualification purposes and uses. The AOs described their own approaches, where they offered a relatively broad range of topics and contexts to begin with and designed their assessments to be fairly 'open', providing options for centres to adapt contexts as appropriate. However, it was also recognised by AOs that topic and task amendment flexibility is mostly used in the context of speaking assessments (both in internal and external models), and that there is less scope or centre resource to adapt the written tests, despite, perhaps, this being in the interest of some learners, which is another reason why some AOs attempt to design assessments that are open. Some AOs allow centres to choose from different assessment versions, to suit individual candidate context.

> At Entry level […] we have always said to centres, if the topic does not suit, tell us, you have to keep the learning outcomes, the assessment criteria have to remain the same and we've provided lots of guidance on how they can adapt, not that they necessarily do because I think it's a lot of hard work, so we try to set assessments that are open enough AO5

> I think maybe more younger learners, […] you would have noticed in our speaking and listening assessments that there are specific role plays for that age group should the assessor find them in the room. AO1
>
> There is a certain amount of adaptation they're allowed to do, but generally speaking, because there are 3 versions available, they can usually pick one that is most appropriate to their learners [rather than adapt]. Granted a lot of them will just take version one and off they go, but there's that possibility there that they can choose which one to use… AO6

Tensions around ensuring comparability between different contexts and topics at the same level, as well as across AOs and over time were also discussed, with AOs acknowledging the challenge but also expressing confidence that they were able to ensure comparability of standards. They also noted that this was mitigated by the clarity of a shared curriculum (as discussed above) which emphasises skills rather than specific context.

> It is still possible because the mark schemes are the same mark schemes, the criteria are the same criteria. So, we can still get that consistency and the comparability across centres, even potentially if a centre may decide to tweak, for example, if they've got to talk about somebody at E1 and that person is slightly tweaked, the same language will come out in describing a person, it may just be a different person that has more relevance to them. […] the issue is about having programmes that are flexible but still rigorous, can still be measured, because obviously what we don't want is something we can't make comparisons between one learner and another, but the way we've developed the qualifications and the assessment do enable them to actually tailor it to different needs. AO5
>
> And I think in the UK system, which is different AOs each offer the same qualification, it does give that standardisation to say that they're all based on the same core curriculum. AO3

While the AOs emphasised the need for flexibility in order to address various learner needs, they also noted that some of the design aspects of these qualifications, or funding patterns, do not quite recognise or cater for certain needs. This was particularly noted in relation to the standard at Entry 1, which was seen by the AOs as not accessible to some learners within the funded guided learning hours, especially those with no basic literacy skills, with some AOs potentially unclear about existing funding arrangements for non-regulated provision.

> […] a lot of people came to us and said they wanted, and it wasn't just the prisons, they wanted pre-entry. The only problem with that is since the demise of the QCF there's no recognition that there is such a thing as something that is actually on a ladder on its way to E1, so we have tried with ESFA for years to get it funded and they would not agree to fund it, but we have got centres, small

> handfuls, who look at it and say this is […] really important with whole groups of learners like asylum seekers and others who have absolutely no language and they're not even, many of them, literate in their own language. So this is really about enabling people to embark on Entry 1. AO5
>
> So, we know at the very low end there are a lot of learners who we probably aren't catering for as well as we'd like to, anybody below Entry 1. AO4
>
> Well, pre-entry skills, pre-entry courses are often not funded, so you might have a pre--entry learner who's put into an Entry 1 course because that's the only way they can access the course. AO1

In related points, some AOs thought that current ESOL SfL provision should also include Level 3, given that some learners need higher language skills for certain employment prospects or progression to further or higher education.

> […] everybody looks at ESOL, it's primarily entry level, but the reality is there are ESOL learners who need to be able to read extended documents and in fact when we developed the ESOL Level 3 it was for a very large FE college with a very large ESOL entry […] and they said they wanted learners to be able to do Level 3, whether it was access to HE, whether it was [access to] Level 3 programmes […] they needed more advanced reading skills and strategies to help them… AO5
>
> So, levels, I think we think levels are really important, but we could probably go lower, and we could probably go higher, I guess, and maybe as we went higher, we could be more specific about the purpose. AO4

Varied assessment models

Perhaps as a direct result of the flexibility currently possible when developing ESOL SfL qualifications, AOs developed and deployed a range of assessment models. However, within that, they perceive the need for the qualifications to be interchangeable between AOs, and comparable in terms of construct and standards.

> […] we work within the framework that all of the other AOs have to work because, and it's really important actually, the principle that these qualifications at the module level are interchangeable is really important for us. […] a learner can pick and mix and combine modules from different qualifications… AO2
>
> I think the diversity between the AOs actually gives flexibility to centres, because, like we said before, the ESOL learners are so diverse that the different methodologies of assessment, it gives them the scope to meet the learners' needs, so that's quite important and at the end of the day everybody is all still focused, they're all being assessed against the core curriculum. So, it is

> standardised, it's just different methodologies of going about getting to the same end goal. AO1

As discussed in the previous section, most apparent was how AOs differed in their use of internal and external assessment, with some AOs using a mixture of both (depending on the skill being assessed and the level of the qualification). The strengths and weaknesses of these approaches were discussed in terms of flexibility, validity and perceived robustness in relation to qualification purposes and use, with an emphasis on the need to provide a balance between ensuring accessibility for all learners and robust controls around the assessments themselves.

> Contextualisation […] is more feasible if you have an internal assessment model, as long as you make sure that it's carried out in the right conditions, as long as we quality assure it, we check for any potential malpractice, make sure it's the learners' own work, all the processes that we have in place. AO5

> So, if you can offer more flexibility, people can actually receive a qualification when they're most ready and they can move on to the next stage of their education more quickly. At the moment we're constrained by the length of a school term. So self-directed learning and all of that is not really supported as well by a fully externally delivered exam; however, if something is completely internal, you have many responsibilities towards the teachers to ensure that they, as evaluators, are fully standardised and trained, and you also have many responsibilities towards the regulatory organisations like Ofqual to ensure that the administration conditions are standardised so that there's been, for instance, no pre-information circulated and people don't have an unfair advantage, well, that the score means what it does. AO2

> The other thing I want to add to that is for centres, a lot of these learners progress internally with them, whether it's on ESOL or onto other qualifications. So, there is very little need or desire for maladministration/malpractice on their part. There is no point a learner passing an Entry 1 qualification if you are going to have to teach them Entry 2 next year. AO6

Some AOs reflected on the benefits of reducing the amount of flexibility at higher levels to recognise potentially higher stakes of those qualifications. Some also suggested that learners at higher levels are more resilient and adaptable so might require less flexibility in assessment and be more able to deal with an unfamiliar assessor or the pressure of an external assessment.

> So, E1 to E3, they could actually break the paper up and take it in sections, so if you've got an E1 learner who has never sat an exam before, they could actually do it in parts, as long as they did it in a secure way and they didn't give them the papers. So, they could also do it in class, which meant they didn't have to have separate exam sessions. […] When we got to level 1, we said at level 1 you do

> the paper as is, in the time required, you sit there, you do it under supervision and then it gets marked, so just like sitting an exam. When it gets to level 2, we felt they needed the formal experience of a proper invigilation. So, it was almost like a gradation. […] And the reason for it was every other level 2, whether it's GCSE English, whether it's functional skills, has always required external assessment and we thought that if people were then going to move on in their careers and were doing level 2, they needed the experience of sitting with invigilation. AO5
>
> […] so, for ESOL reading, we thought that was achievable and moved that component at level 1 and 2 onto an externally set and marked model. AO4

There were also suggestions that some of the assessment approaches reflected the ethos of the AOs, or their wider understanding of the construct of assessment in related areas, reflecting their 'brand' in that respect.

> […] the core curriculum and the other frameworks that have also informed the design of the test are important, but I think we've also referenced general English language assessment principles and [our] ethos, our assessment expertise and those things also inform the final shape of the qualification, so it's not just a paint by numbers exercise. AO2
>
> So, based on the standards, linked to the Core Curriculum and I think we've always gone for as much internal assessment with internal quality assurance as possible, because that's always been more of a manageable approach for a lot of our centres really and our learners, with external quality assurance on top. AO4

Another prominent area where assessment models differed between the AOs was in relation to how they conceptualised their measurement models and pass grade standards. As already noted, some AOs implement a compensatory model, while the majority implement varied conceptions of mastery. Within the mastery models, there were differences between AOs with regard to how they conceptualised mastery and how 'hurdles' at AC, task or question level were used in their assessments. This partly reflects differing interpretation of the core curriculum and the QCF, as well as differences between AOs in their assessment philosophies and preferred strategies. As previously discussed, some AOs that employ mastery models acknowledged that this was a historical QCF-motivated approach and thought that this did not necessarily represent the best practice approach to summative assessment of language acquisition. Several AOs pointed out that defining measurement models and/or what represents sufficient evidence of attainment is a non-trivial task even where there are attempts to come to decisions through discussion of a range of stakeholders.

> And that may work for vocational qualifications, but at the time of having to move from the NQF to the QCF, this was something that we discussed a lot and that actually maybe language learning or literacy is not really a series of boxes to be

> ticked. […] and this was discussed at considerable length. […] That's the nature of QCF, if you don't know your alphabetical order you can't achieve the qualification and therefore what we tried to do was to make sure that there were equal opportunities to achieve each of those assessment criteria. AO3
>
> So, I would hate to be hung, drawn and quartered over my inability to use a semicolon properly or whether or not an Oxford comma was the right thing to use or not, for instance. So, since that would be one of the minimum criteria, I would be very uncomfortable with using a mastery model unless I had complete confidence in the descriptions of mastery. […] But also, may I add, this is the eternal problem, how much is enough, when is someone qualified to pass, so regardless of the model the question will always be there. AO2
>
> From an assessment perspective I think it makes it very difficult in the standard-setting process, that mastery, because you're almost balancing off 2 things against each other, because you're saying, you need to be able to do all these things, but then you're setting a pass mark and saying well it's OK not to meet all those things. I also don't necessarily think that language works in that way. I don't think it's the same as competency-based skills type of assessment that you might see within an NVQ or something like that, I think it's a lot easier to grade and to sit learners on a scale, so, I, from an assessment perspective, think it would be helpful if there wasn't that mastery component. AO4

Some AOs discussed the use of task- or AC-level hurdles in relation to a higher level of 'strictness' of SfL qualification subject-level criteria compared to qualifications such as GCSE which use compensatory models. Some AOs explained that task-level hurdles were there to ensure that the full range of learning objectives is covered, while others explained the use of both AC-level hurdles and numerical pass mark in reading comprehension as intended to ensure wider coverage of the domain rather than evidence a minimum mastery level. There was also an indication from some AOs that hurdles may not always be an optimal approach but can be used to ensure coverage of required content areas.

> […] the qualification subject criteria is quite strict as well for ESOL […] with our papers there's section 1, section 2, section 3, you've got to pass each section to pass the overall qualification, so, again, in that respect it's much tighter, it's much stricter than for GCSE or even functional skills, for example, where you've just got to, holistically you've got to pass the paper. There are hurdles within ESOL where you have to be able to demonstrate the skills to a certain level, even within the subskills. AO6
>
> […] for writing they have to pass every task and that was to ensure there was coverage of all the different learning outcomes and the learning outcomes and assessment criteria. AO5

> […] And I think that comes from when the qualification was being developed. It was felt that a numerical mark should be included on top of the assessment criteria requirement. Because particularly for reading it was felt that learners needed to be able to demonstrate that kind of understanding of breadth and range of lexis AO3
>
> […] the way the mark scheme was structured would enable people to pass [without getting many marks for spelling, punctuation and grammar], so we looked at it and we decided that the quickest way to do it was to actually say OK, and I'm not a great fan of hurdles, but there is a hurdle. AO5

Valued qualifications but room for change

AOs, perhaps unsurprisingly, saw these qualifications as having enormous value to learners and society as a whole. They also largely thought that they had appropriate processes in place for their development, and that their assessments were high quality and valid.

However, there was a sense from all AOs that these qualifications were undervalued in terms of funding, and not sufficiently recognised for their potential to enable learner progression, for instance to apprenticeships, or T Levels, particularly in comparison to Functional Skills English qualifications, and that this affected stakeholder attitudes towards them, including centres, employers, higher education institutions, and learners themselves.

This was seen as unjustified, and all AOs emphasised the need to recognise parity between these qualifications and similar qualifications such as Functional Skills or GCSE English, as well as some ESOL International qualifications. The SfL qualifications were not seen as dissimilar in their enabling function to ESOL International qualifications, some of which are widely recognised, at competence levels similar to L1 or L2, by reputable higher education institutions.

> If ESOL ends up being a poor relation that is not recognised in the same way as functional skills then you're not serving the employers or the higher education and you're not serving the learners, because what you're doing is you're giving them less competence to do what they want to do by leaving out the ESOL elements and encouraging them, because of the funding regime, to take functional skills because it's funded and then they're not as skilled as they need to be to do what they want to be. AO3
>
> I think funding also impacts, that's a barrier for learners for all levels, because they can't access the funding or centres can't access the funding and there are vast swathes of people that would benefit from ESOL Skills for Life that can't actually access it. AO1

> The brand of ESOL… there's an awful lot of work's gone into the brand of functional skills that hasn't gone into the brand of ESOL and they do have very different purposes… […] There has been a bit of a missed opportunity over time, because ESOL hasn't been accepted as evidence of language capability for apprenticeships and so there's a group of learners there who often end up doing functional skills because they can't access maybe GCSE in the timeframes that they've got. So, they opt for functional skills to meet the maths and English requirements for apprenticeships, but actually that ESOL qualification would be a really good piece of evidence. […] There's huge potential for this qualification in things like T Levels. There's a transition year, there's an opportunity for people – there's a missed opportunity actually for ESOL learners potentially to make the transition to T Levels, because there's going to be an expectation for one or two named qualifications for eligibility. Why not ESOL? It's the most challenging qualification at Level 2 and […] often learners come from educational backgrounds which are quite strong, just lack the language. AO4

> If you want to go and study at what you would class as a decent university then you're not getting into that university with an ESOL qualification, you would have to do a Pearson Test of English or IELTS or get a GCSE, so you have to question the purpose of ESOL in respect to that and in respect to younger people. […] I think the other key thing is then the marketing of these qualifications and the way that they're marketed in general to promote the benefits of these qualifications, firstly, to learners and, secondly, to employers and universities and stuff like that. […] So, again, that is something that really needs to be addressed, because all we're doing really in the entirety of all this is saying that ESOL just isn't quite as good as GCSE or functional skills and if that is what we're saying, then why would anybody want to do ESOL when you get to the higher levels, why would you want to spend 270 hours doing a qualification that's just enough to make you get onto another qualification at the same level? […] and I think that's the fundamental issue is how ESOL is positioned is, it's not the qualification itself, it's not the core curriculum, it's the simple fact is it is an afterthought in government policy. AO6

It is also important to note that the COVID-19 pandemic has had an effect on the way that ESOL SfL qualifications are currently being assessed and this may have a lasting legacy in terms of how certain assessment methods are delivered. Several AOs mentioned that they were thinking about, or in the process of developing, online assessments or ways to support centres with online delivery of speaking assessments.

> I think that most of our centres feel that in-centre assessment is really important and the most supportive way to deliver these qualifications, because of the very specific nature of the candidates and the support that they need. However, they

> do see some value in online qualifications as well, online adaptations as well. AO3
>
> I think there are some provisions also that are setting up with, because it's virtual they're actually recruiting learners from all over the place, so they actually haven't got a centre, they are now an online provider. So that's, the way forward is they need online assessments. AO1

AOs also highlighted a range of barriers to SfL qualification and assessment delivery during Covid that are related to the nature of the cohort, particularly technological barriers, and particularly at entry level. On the other hand, AOs mentioned resourcefulness and perseverance of ESOL teachers, who continued to come up with digital resources to support their learners as far as possible.

> I think we've learnt an awful lot about who our ESOL learners are and the challenges they face, particularly at entry level I would say, even more so than level 1 and level 2, like the refugees and people like that that we've talked about earlier on, who have already faced so much adversity and then all they want is their certificate to show what they've done, and they can't get there because they've not got the technology to support it. […] It was really fascinating and actually the ESOL teachers, very creative, finding ways of using apps and different things and there is a richness in EFL actually that a lot of them were drawing from, because there is a history there of quite a lot of remote teaching and learning. AO4
>
> I think one of the other issues going on at the moment is how COVID has adversely affected the communities that are traditionally ESOL in their nature, probably more so than other communities as well and also looking at the delivery models within colleges and providers when we're talking about a community approach, a lot of these community places are not available at this moment in time. AO6

As previously discussed, there was an acknowledgement that there was room for refinement in the core curriculum and that further efforts could also be made to improve consistency across AOs as well as reconsidering some of the current, sometimes legacy, assessment strategies. However, AOs were keen that any reform or tightening of rules should not come at the cost of flexibility to the learners, a vital element of what makes ESOL SfL qualifications so valuable to those that take them.

> There are different ways of regulating a qualification, but I would really say anything that creates more barriers to participation or creates more barriers to achievement, by removing flexibility, should be absolutely off the agenda. AO3
>
> So, there was perhaps a hope, a preference to keep things standardised all the way from A levels and GCSEs through to the different VTQs without a full

acknowledgement of the nuanced differences between these different quals. That would be my big thing and it puts enormous pressure on AOs if they have to fit into a set of controls that don't quite work for that qual and there's limited ability to negotiate. AO2

[…] there's a fine line, isn't there, you want guidance, but you don't want something that dictates and becomes completely inflexible. I think we would welcome the opportunity for more consistency across the products, but I don't think it needs to go as far as what functional skills have gone. AO4

Study 3: Curriculum alignment review

Introduction

Alongside Ofqual GCR, the only criterion for the AOs in relation to SfL qualifications is that they meet the requirements of the NSAL and demonstrate a clear relationship to ESOL core curriculum (which also reflects the NSAL). From our assessment review we could see that there were some differences in approaches to assessment and interpretations of the curriculum content statements between AOs, differences in weighting of assessment criteria related to specific content statements, differences in descriptions of pass grade standards, and so on. While different approaches to assessing certain constructs may be equally valid, a wide range of divergence between AOs might also lead to a divergence in standards at individual component as well as qualification level. We conducted this curriculum alignment review to enhance our understanding of the extent to which different AO approaches substantively align with the NSAL and the curriculum.

Method

Two SfL/ESOL subject experts were recruited to conduct this review. Both reviewers also participated in our study 4, which preceded the curriculum alignment review.

One of the reviewers had extensive expertise in teaching, teacher training, curriculum training and examining of SfL qualifications (as well as their predecessors and other ESOL qualifications) and was thus thoroughly familiar with the detail of the core curriculum. This reviewer also had extensive background and understanding of the history of the core curriculum and SfL qualifications development, as well as a detailed understanding of related policies and funding arrangements, and has held a number of senior positions in the ESOL area in several institutions within the further education sector. The other reviewer had a more extensive background in ESOL International and EFL qualifications as well as CEFR expertise primarily as teacher and assessor, and past experience of teaching MFL and adult literacy qualifications and of qualification development. This reviewer also carried out the CEFR content mapping for our Study 4 ahead of the curriculum alignment review, and therefore additionally gained familiarity with SfL specifications and the curriculum through that exercise. Both reviewers also had a research or lecturing background in linguistics as well as broader educational research, and experience of working as Ofsted inspectors in the further education and skills sector including ESOL SfL.

Neither reviewer had specific experience of operational assessment design from any work for awarding organisations though had clearly accumulated reasonable

expertise through their teaching and examining as well as consultancy work developing MFL or teacher training qualifications. This is a potential limitation of our choice of reviewers. However, this study was primarily concerned with curriculum alignment, and, while some questions were asked about general assessment quality too, we interpreted subject expert responses to those questions alongside our own internal assessment review, which was carried out by an experienced qualification and assessment developer. We also sought views regarding assessment properties of the SfL qualifications from several participants from our Study 4, who had extensive assessment development expertise from working with a range of testing organisations delivering ESOL International or EFL qualifications.

Overall, given that it is sometimes challenging to identify suitably well rounded subject experts for these kinds of analyses, we believe that our reviewers had appropriate expertise and understanding of the key issues to carry out a thorough and authoritative curriculum review. While we cannot discount the possibility that their individual views were in some ways skewed or biased, this should have been to some extent reduced through discussion in order to agree on key points, which was part of the study design. Ideally, of course, views from a wider pool of experts might well be worth seeking were there to be a reform programme for these qualifications or the curriculum in the future.

The subject experts conducted the review against a series of questions that we asked in relation to each of the 4 AOs in our sample, and for each of Entry 1, Entry 3 and Level 1 SfL components[10]. For each AO and level, the reviewers considered qualification specification documents and any accompanying documents (such as mark schemes, assessor guidance) and 2 sets of live assessment materials for each component, provided to us by the AOs.

Ahead of starting the review, a briefing meeting was held with each of the experts to discuss the questions that they would be asked to consider during the review and clarify the requirements. The experts carried out the review independently of each other. They were asked to provide justification and examples for their views and observations.

The questions (see Appendix 2) focused around 4 key areas. They asked whether each AO's SfL assessments:

- target the standards as appropriate to the relevant curriculum level
- assess an appropriate sample of the curriculum

[10] Even though speaking and listening form a single component in SfL qualifications, listening comprehension was analysed separately in this study. Listening also has a separate set of standards and section dedicated to it in the curriculum at each level.

- interpret the curriculum appropriately, and, where they might not, whether this is attributable to ambiguities in the curriculum itself
- assess the content effectively and validly

Across these main areas, there were 12 questions that the subject experts answered for each component (reading, writing, speaking and listening) and 1 question that related just to writing and speaking. In addition to these AO-specific questions, the subject experts also answered questions about:

- comparability of approaches, demands and pass criteria between the AOs at each level
- whether there are aspects of the curriculum that might benefit from updating or amending at each level
- whether it would be beneficial to align the curriculum (and SfL qualifications) more explicitly with the CEFR

Subject expert responses were analysed by an Ofqual researcher. Most of the analysis was qualitative, and essentially summarises the expert views on our different questions, identifying the main themes that the experts raised. In addition, in order to get a sense of general patterns across different components and levels, the responses were coded in terms of whether the relevant AO/component was deemed fully (Y), mostly (M), partly (P) or not appropriate (N) in relation to each question. At each level, this resulted in 50 ratings for each of the 4 AOs (12 for each of reading and listening; 13 for each of writing and speaking).

Due to limited resource, no double coding was undertaken, which represents a potential limitation of this analysis. However, a discussion meeting was held with the subject experts to seek clarification for their responses where necessary (for instance, where their view was unclear and could not be coded using one of the codes above) and engage them in discussion to resolve any significant points of disagreement between the 2 experts in relation to specific questions. The codes were updated following the meeting where necessary.

The results section below summarises the main points made by the reviewers and findings based on the patterns of ratings across AOs, components and levels, highlighting key areas of concern and pointing out areas where curriculum alignment could be improved.

Results

Summary of findings

As can be seen from the tables below, summarising across the AOs and questions, there is a reasonable proportion of categories (Y and M) indicating that the assessments were seen as appropriate in relation to the questions we were asking and where the assessments are well aligned with the curriculum or reflect good assessment practice in relation to the curriculum. Speaking was generally seen as more appropriately aligned and assessed compared to the other components. However, a range of alignment and other issues were also raised, and the analysis in this section primarily focuses on those as potential areas for improvement.

Table 1 Proportions of ratings characterising assessments as fully (Y), mostly (M), partly (P) or not appropriate (N) – Entry 1

Category	Reading	Writing	Speaking	Listening	Across components
Y%	40	40	65	60	52
M%	33	37	21	13	26
P%	10	15	13	27	17
N%	17	8	0	0	6

Table 2 Proportions of ratings characterising assessments as fully (Y), mostly (M), partly (P) or not appropriate (N) – Entry 3

Category	Reading	Writing	Speaking	Listening	Across components
Y%	33	42	58	35	43
M%	19	17	8	10	14
P%	27	25	25	48	31
N%	21	15	10	6	13

Table 3 Proportions of ratings characterising assessments as fully (Y), mostly (M), partly (P) or not appropriate (N) – Level 1

Category	Reading	Writing	Speaking	Listening	Across components
Y%	40	38	56	31	42
M%	13	13	13	10	13
P%	29	37	25	48	35
N%	19	12	6	10	12

It is notable that the overall appropriateness of curriculum alignment of the assessments seems to decline from Entry 1 to Level 1 for most of the AOs in our sample, based on the increasing proportion of categories characterising assessments as partly appropriate (P) or not appropriate (N). This is concerning given that the stakes of the qualifications are likely to increase at higher levels, as these are the levels that should allow progression to skilled employment or further education most directly.

With some exceptions, the demand of assessments and pass grade standards at Entry 1 was seen as broadly appropriate in relation to the standards implied in the curriculum. Where the demand at Entry 1 was not deemed appropriate, it was characterised as too high for that level. At Entry 3 and Level 1, the demand was considered less appropriate, and generally too low.

In terms of curriculum coverage, the review confirmed that the specifications reference most of the curriculum statements. However, this was sometimes deemed not to be at an appropriate level, with insufficient sampling across the range of contexts and functions specified in the curriculum, particularly at Entry 3 and Level 1. For instance, there is an emphasis on assessing form filling at every level in both reading and writing, even though they are just one example of many text types that could be required from candidates.

In some cases, the reviewers thought that the curriculum statements are assessed in the way that demonstrates adherence to the letter but not sufficiently the spirit of the curriculum. For instance, there seems to be insufficient focus on reading for meaning at Entry 3 or reading for detail at Level 1, with subskills being assessed in isolation rather than as components contributing to and enabling the global skill of reading comprehension. Similarly, certain enabling skills such as knowledge of alphabetical ordering at Entry 3 or declarative knowledge of text and grammar features at Entry 3 and partly Level 1 appear to be over-emphasised in the assessments, contrary to the apparent intention of the curriculum, which aims at developing the understanding of the meaning of these features rather than labelling them. These content statements are sometimes given ‘hurdle’ status in some assessments, where a pass grade could not be achieved without achieving those criteria/answering those questions correctly. As another example, at Level 1, the NSAL standard “trace and understand the main events of continuous descriptive, explanatory and persuasive texts” was deemed to have been taken too literally to mean chronological events and led to one long text per test often being chosen to have dates in it, which was deemed an unnecessary constraint and not in the spirit of the relevant standard, which also includes non-chronological examples.

In most cases it was not clear why such apparent misinterpretation or narrowing may have occurred, as the curriculum statements and examples were seen by the reviewers as sufficiently clear and indicative of the appropriate level and range of contexts and functions, which the curriculum lays out in terms of increasing complexity through the levels. It may be important, however, to confirm with other experts and relevant AOs whether there is indeed sufficient clarity about this in the curriculum and understand any additional constraints that may have driven some of the narrowing of the contexts and functions.

It was also highlighted, however, that there is sometimes a difficult balance to be struck between ensuring the breadth of coverage in assessments while not

jeopardising depth of coverage in teaching. In addition, the atomistic nature of the ACs and the apparent requirement for exhaustive sampling of the curriculum in assessment in the mastery model were seen as partly to blame for some of the above-mentioned issues.

According to the reviewers, the assessments appear to offer some room for improvement, to varying degrees depending on level, AO and component, in relation to:

- reading comprehension – item quality and plausibility, better targeting of item demands for the level, between-version comparability, text variety and authenticity
- writing – better targeting of item/task demands for the level, topic comparability between versions, prompt quality/amount, assessor guidance, marking criteria clarity, assessment approach consistency between AOs
- speaking – topic comparability between versions, prompt quality, assessor guidance, marking criteria clarity, assessment approach consistency between AOs
- listening comprehension – item quality, better targeting of item/task demands for the level, assessment approach consistency between AOs (use of audio or not, written questions or not), audio speed, variety/range of audio input

Comparability between AOs across reading components at each level in terms of stimulus text demands or pass grade standards was deemed insufficient. In relation to writing and speaking, in some cases it was difficult for the reviewers to make a judgement about comparability of demand and pass grade standards given the differences in assessment approaches, weighting of criteria and sometimes lack of clear marking criteria. However, the range of differences, as well as lack of clear marking criteria, particularly in internal assessment models, were in themselves suggested as involving threats to comparability.

The curriculum itself was deemed to be largely fit for purpose. The key issue seems to be finding a way to clarify curriculum messages and standards amongst users, including the AOs. As an example, the focus on more academic and professional reading and writing at higher levels, which is largely missing from the assessments, is present in the curriculum. It is not clear why it does not feature sufficiently in the assessments, given that some of the purposes of SfL qualifications, especially at Level 1 and above, could involve progression to further study and employment. The reviewers pointed out that there are resources in the curriculum itself, as well as other resources, which could help with pitching the assessments at the appropriate curriculum level.

The reviewers thought that the curriculum itself is largely consistent with the CEFR in terms of the language competencies it involves and should therefore allow alignment

of the SfL qualifications to the CEFR. Overall, the reviewers considered it potentially beneficial to map these qualifications to the CEFR. This was deemed as potentially helpful for increasing recognition of SfL qualifications (for instance, for higher education progression or for comparison with ESOL International qualifications), as well as promoting and clarifying alignment between AOs (although appropriate mapping to the curriculum would also enable that).

Below we provide a more detailed analysis of the findings from the review, grouped into several broad themes.

Assessment demand and pass grade standards

At Entry 1, the demand of assessments was seen as broadly appropriate according to NSAL and the curriculum standards. However, there were some examples of standards that are potentially too high for the level in some reading and listening assessments, as well as placing too much weight on assessment criteria related to spelling and punctuation, which is not required by the curriculum at Entry 1. Pass grade standards, alongside the coverage of the content domain, were deemed to provide sufficient evidence of mastery for most AOs and components at this level, except for one AO's reading comprehension, and another's writing component.

At Entry 3 and Level 1, the demand was considered less appropriate, and generally too low. At these levels, only one AO's assessment was deemed entirely or mostly appropriate across most components (although potentially too high in some respects for writing and speaking, particularly in relation to the breadth of topics and length of assessment). In other cases, some component demands were deemed appropriate while others were not and were typically deemed as more appropriate in speaking than in reading or writing components, depending on the AO in question.

In terms of pass grade standards at Entry 3 and Level 1, again, only one AO was consistently deemed to evidence mastery of the curriculum at the relevant level. The others were deemed not to provide sufficient evidence of mastery, particularly in reading comprehension and writing, and in some cases speaking and listening comprehension. The reviewers pointed out that, in some cases, unclear marking criteria made it difficult to establish what the pass grade standard entails, and that, to establish this, it would be necessary to consider samples of writing and speaking at the pass grade boundary (as was done in our Study 4).

Across writing and speaking, the reviewers found the marking criteria often less than clear, which is particularly concerning where assessment is internal. It would be important for validity and transparency that the marking criteria are made clearer, particularly with respect to ranges of required language (for instance, vocabulary, grammar, functions) at each level, as well as perhaps allowing markers to distinguish

qualitatively between performance levels rather than just determining whether certain ACs are met or not met.

Curriculum coverage

In terms of curriculum coverage, the review confirmed that the specifications reference most of the curriculum statements, although there was evidence of some narrowing of the curriculum, with insufficient sampling across the range of contexts and functions specified. In some cases, curriculum statements are assessed in the way that demonstrates adherence to the letter but not sufficiently the spirit of the curriculum according to the reviewers, with issues noted in relation to interpretation of certain aspects of the curriculum. This was observed mostly at Entry 3 and Level 1.

Particular issues were noted with the reading comprehension component, with not enough focus on reading for meaning at Entry 3 or reading for detail at Level 1, with subskills being assessed in isolation rather than as components contributing to and enabling the global skill of reading comprehension. At Level 1, it was noted that there is little focus on reading for inference, opinion and details that require a more in-depth understanding. In addition, reviewers thought that there was not enough coverage of more formal aspects of language (for example, formal reports in reading at Level 1, more formal exchanges or more formal audio material at both Entry 3 and Level 1 in speaking and listening, more complex topics/types such as social problems, environment or reports in Level 1 writing) even though these are present and exemplified in the curriculum and would suit the apparent purposes of these qualifications, especially at Level 1 and above. There are also inconsistencies between AOs in the amount of interaction assessed through writing and speaking, with one AO assessing these aspects to a significantly lesser degree than the others.

The reviewers also noted instances where assessments appear to over-emphasise or over-assess certain aspects of the curriculum. These are often enabling subskills, which, according to the reviewers, should not be assessed in isolation or for their own sake, and which are singled out in the curriculum for instruction rather than assessment purposes (for instance, assessing alphabet and alphabetical ordering at Entry 3 or assessing knowledge of names of grammatical and text features rather than how their function interacts with comprehension of text in both Entry 3 and Level 1 reading comprehension). Such approach to assessment of these aspects was deemed not to be in the spirit of the curriculum, which aims at developing the understanding of the meaning of these features rather than labelling them.

At Entry 1, too much weight in assessment seems to be placed on criteria related to spelling and punctuation, which is not required by the curriculum at this level. There also appears to be too much focus on assessing simple form filling at each level, both as part of reading comprehension and in writing. At Level 1 specifically,

although forms are mentioned in the curriculum, this is only in connection with complex open responses, which are not the focus in the assessments, and they are one example of many text types that could be required from candidates.

Important inconsistencies were noted in relation to listening comprehension. At Entry 1, it was noted that the curriculum requires assessment of listening comprehension only in face-to-face communication, rather than listening to audio recordings or conversations between other people. Nevertheless, most AOs do include listening to audio recordings and answering specific comprehension questions in relation to those. This was deemed as potentially problematic, and reviewers thought that the recordings tend to be too information-heavy for Entry 1.

At higher levels, it was noted that the curriculum requires listening comprehension of audio recordings as well as in face-to-face communication. Most, though not all, AOs do include assessment of audio comprehension at higher levels. However, where this is included, the reviewers commented that the comprehension tasks are typically limited to very few, often fairly undemanding and sometimes poorly drafted questions, and the audio recordings are sometimes too slow for the level. At Level 1 in particular, NSAL focus on ‘formal exchanges connected with education, training, work and social roles’ or listening to authentic material on radio, TV. This type of audio or this level of topic complexity was not apparent in most of the reviewed listening assessment materials at Level 1, with most assessments under-representing the domain in this respect.

Assessment approaches

Reading comprehension

With regard to reading comprehension, at Entry 1, the approaches were seen as broadly appropriate, although reading load of the rubrics and items and some text demand/lengths were deemed too high in some cases. Some of the questions require reading for detail, meta-language skills such as naming text purposes, understanding of anaphoric pronoun reference and use of capitalisation rules that are too demanding for Entry 1. Some items were considered not entirely construct relevant, too guessable and sometimes involving implausible distractors.

At Entry 3, some form-filling tasks in reading assessments were deemed too low in demand, with most vocabulary at Entry 1 and the form features at Entry 2. There were instances of Entry 1 or Entry 2 level questions in relation to other tasks too. In some cases, the reading stimulus texts were considered to be too low level for Entry 3, and sometimes did not involve significant paragraphs, which are required at this level. Approaches of assessing within a single topic area were not considered good

practice and were deemed likely to limit the range of vocabulary that the candidates should demonstrate at this level.

At Level 1, more formal course- or work-related texts are not given enough focus according to the reviewers. They noted that the curriculum is pushing towards more complex reading at this level and looking at how meaning is built up over several paragraphs, as preparation for vocational and academic courses. Some stimulus texts do not involve long enough paragraphs for the level and thus do not enable sufficient assessment of the above-mentioned comprehension skill. The approach of assessing within a single topic was considered even less appropriate at Level 1.

Across Entry 3 and Level 1, a number of reading comprehension items were deemed poor in terms of demand targeting to the level. Reviewers commented that there were too many true/false items in some assessments, which are guessable, as well as a number of items with implausible distractors or items giving answers to other items in the test.

Writing

Although most writing tasks were deemed appropriate across AOs and levels, some were deemed insufficiently authentic and sometimes not designed sufficiently well to elicit the required language from candidates. In some cases, this was because of insufficient prompts, which could affect learner performance. In other cases, especially at Entry 1, some tasks appear to require grammar that is too high for the level, for instance assessment of present continuous, or a word count that is too high for that level.

At Entry 3, some word counts were deemed insufficient for the candidates to demonstrate production of some important features of writing at Entry 3 such as development of ideas through paragraphing. Some writing tasks involve very simple form filling, which was considered to be below Entry 3. Some tasks were not deemed sufficiently communicative or suitable for assessing written production. On the other hand, while broadly compatible with the demands of the curriculum, some writing assessments were considered too demanding in the number of diverse pieces of writing required.

There are similar issues with the writing components at Level 1. In addition, according to the reviewers, the writing tasks sometimes do not provide enough range of audience and purpose for this level and have little focus on more complex topics/types such as social problems, environment, or reports.

Speaking and Listening

With regard to listening comprehension, as already noted, most AOs include audio comprehension tasks at each level. However, the reviewers characterised these as

involving few sometimes undemanding and poorly drafted questions, while the audio recordings were sometimes deemed to be too slow for the level at Entry 3 and Level 1, and too information-heavy for Entry 1. The AOs which include audio recordings in their assessment also assess aspects of listening comprehension through their speaking assessments. Lack of use of audio recordings in comprehension tasks at higher levels was not considered appropriate.

Overall, in terms of approaches, speaking was deemed to be assessed best across most AOs, as well as being mostly appropriate in matching the spirit of the curriculum. Again, however, the reviewers thought that some AOs could improve their prompts to ensure that the required language is elicited from the candidates (particularly where assessment is internal) and include more interactive tasks.

Comparability of demand

There was evidence of lack of comparability within AOs (between different assessment versions within components), as well as between AOs. Comparability of demand between AOs across reading components at each level was seen as insufficient. Issues were observed in relation to comparability of pass grade requirements in particular, but also vocabulary demand in texts and questions, amount of reading time and extent of requirements for detailed comprehension, text complexity in terms of number and length of paragraphs, and item quality.

In relation to writing and speaking, in some cases it was difficult for the reviewers to make a judgement about comparability of demand and pass grade standards given the differences in weighting of criteria and assessment approaches as well as sometimes unclear marking criteria.

The range of differences, as well as unclear marking criteria, particularly in internal assessment models, were in themselves suggested as constituting threats to comparability. Some of the comparability issues noted in writing include varying word count ranges, variable time allowance, variable extent of requirement for extended writing at Entry 3 and Level 1, varying requirements and opportunities for use of complex sentences and discourse markers, and varying spelling and punctuation accuracy criteria. In speaking, there were comparability issues including varying requirements for length of speech, varying extent of use and nature of interactive tasks, varying emphasis on the requirements to display a range of grammar and functions in production, and varying amount of task prompts and administrative guidance.

Atomistic assessment criteria and the mastery model

The atomistic nature of ACs and apparent exhaustive sampling of the curriculum was deemed partly to blame for inordinate weight being given to some enabling subskills,

which, according to the reviewers, should not be assessed in isolation or for their own sake, and which are singled out in the curriculum for instruction rather than assessment purposes. The reviewers suggested that the curriculum is sufficiently clear regarding this and implies teaching and assessment of increasingly complex or wide range of skills and contexts through the levels. It may be important to confirm with other experts or AOs whether there is indeed sufficient clarity about this in the curriculum.

This was also thought to lead to an inappropriately wide range of question demands in reading assessments, where questions on basic enabling skills which are taught at the start of Entry 1 for instance, sit alongside much more demanding questions targeting much more complex skills, or even skills that are beyond Entry 1:

> The component skills range in level from the beginning of E1 ('understand that print carries meaning') to the end of E1, and AOs are not distinguishing between these 2 ends of the scale sufficiently, so some very basic tasks and other much harder ones are included. Recognising text type at this level is intended as a development activity to support understanding of a text, not as a skill in itself. (R2)

The mastery measurement model, where each criterion has to be met a minimum number of times or to a minimum degree, was not seen as particularly helpful or necessary to ensure appropriate performance standards. In the words of the reviewers:

> There is a danger with a competence-based approach of "not seeing the wood for the trees"; language is much more than the sum of its parts. On occasion the mark schemes produce a pass because the candidate has ticked off each criterion at least once, without some kind of holistic overview of whether there has been genuine communication and task fulfilment. … Where it works best is where there is a minimum mark as well as criteria coverage, and where […] an overview is taken. (R1)

> The criterion-based approach causes difficulties in assessing levels unless very clear guidance is given to markers. Unless the bar is set very low, compensatory approach may be fairer for candidates too, as writing covers a range of subskills and not everyone has the same strengths and difficulties but can still achieve communication. (R2)

Indeed, previous research in the context of internal assessment within the CBA tradition suggests that despite atomisation of assessment criteria and use of mastery model, there is evidence that assessors still make holistic, contextualised and compensatory assessment decisions (Torrance et al, 2005; Johnson, 2008; Carter & Bathmaker, 2017), in some cases disregarding assessment criteria when it may disadvantage a student they believe is 'worth' a certain grade (Garland, 1998). It

may, therefore, be important to reconsider the suitability of the mastery model for assessing language competence as implemented in current SfL qualifications.

CEFR alignment

Overall, the reviewers considered it potentially beneficial to map these qualifications to the CEFR. This was deemed as potentially helpful for increasing their recognition, as well as promoting and clarifying alignment between AOs (although they thought that appropriate mapping to the curriculum should also enable that). The reviewers thought that the curriculum itself is largely consistent with the CEFR in terms of the language competencies it involves and should therefore allow alignment of the qualifications based on the curriculum to the CEFR.

> It would be helpful to link the qualification officially to the CEFR like other ESOL qualifications. This would allow comparison with other qualifications, for HE requirement purposes, and also international recognition where appropriate. Using RQF levels and measuring achievement against that of English first language speakers is both unhelpful and misleading. […] SfL Entry 3 is comparable with a high GCSE (Level 2) in MFL. It would make much more sense to align with the CEFR and foreign language achievement. (R1)
>
> Looking again at A1 of CEFR, I think the ECC aligns pretty well, as long as you ignore some of the literacy-related enabling skills that CEFR doesn't mention (but does assume). However, as we have seen, some of the E1 tests contain quite a lot that is (wrongly) at A2, as well as some things (eg, signs) that are pre-A1. This realignment to A1 would help those who want a more basic qual than Entry 1 – as it would be more basic for reading than now (with some AOs). S&L and writing are better aligned to A1, though some AOs' listening tests are too high for A1 and some writing is too demanding for A1 (and Entry 1). (R2)
>
> Yes – it would be helpful, as ESOL core curriculum does align Entry 3 to B1 MID at least, B1+ in places. Otherwise, the jump to Level 1 will continue to be too great for students. (R2)

Improving the understanding of the curriculum messages

The review suggests some scope for updating the curriculum in terms of adding more reference to some more up-to-date topics. However, this was seen as minor, and the curriculum was perceived to be largely fit for purpose. The key issue seems to be finding a way to clarify curriculum messages and standards to users, including the AOs:

> I think the problem is not the curriculum but the way it is used by dividing into separate criteria for assessment. The 'component skills' are intended for teaching and not for assessing everyone, eg, no AOs assess phonics as this is too difficult in a test and anyway is something for teaching, as matching upper and lower case and recognising digits should be (whereas writing upper and lower case and digits is OK to assess). Recognising text type at this level [Entry 1] is intended as a development activity to support understanding of a text, not as a skill in itself. (R2)

The reviewers pointed out that there are resources in the curriculum itself that do not seem to be sufficiently used by the AOs, which could help with pitching the assessments to the appropriate curriculum level:

> However, the useful list of the communicative functions expected at this level [Level 1] (at the end of the level in the Core Curriculum) gets insufficient prominence in the assessments and could ensure a more exacting assessment. […] It includes describing, comparing, narrating in past, obligation (and absence of – "you must/don't have to"), reporting, requests, hypothesise, explaining and giving reasons, feelings, likes, dislikes, opinions, wishes and hopes, advice and suggestions, recommendations, persuading, complaining, warning. This provides a much clearer sense of what should be expected than the standards alone. (R1)

> Some AOs have misinterpreted the level, so it seems there aren't enough examples of the level. But ESOL Exemplars[11] should help. They could also refer to SfL learner materials, which are quite accurate on levels too and still available on Excellence Gateway. There are comprehensive lists of grammar and functions in the curriculum and it seems that AOs are not referring to these. At the very least they should all reprint these in their guidance to assessors, or better, elaborate on them [as some AOs do], to encourage assessors to use them more. (R2)

[11] ESOL Exemplars (DfES, 2004) was a resource developed for teachers and trainers, commissioned by the Adult Basic Skills Strategy Unit and supported by an advisory group drawn from ESOL awarding organisations and QCA consultants. They were meant to support establishment and growth of a common understanding of what constitutes progression in ESOL, in teaching, learning and assessment. They contain learners' speech, writing and reading texts, with comments on relevant aspects of language used and linked to the NSAL. The aim was to aid users in understanding 'levelness', the level of learners' language at different levels of NSAL and were intended to be aligned with the SfL awards at the time. They show the work of learners who are mainly secure in a level, but also have skills which may be below or above that level.

Study 4: Comparability of standards and CEFR alignment

Introduction

The purpose of this study was twofold. One was to compare assessment and performance standards at the pass grade between the SfL qualifications in our sample to determine if there are any notable discrepancies. As demonstrated previously in this report, there appears to be a range of potential threats to comparability between different SfL qualifications arising from different assessment approaches or interpretations of the curriculum. A direct comparison of pass grade standards should provide a better indication of whether the threats to comparability are actually realised in practice. Notable differences in grading standards between similar qualifications at the same level might stand in the way of learner progression or their ability to use the qualifications from different AOs interchangeably, thus potentially reducing the flexibility in SfL provision.

The other purpose was to align aspects of these SfL qualifications to the CEFR. The CEFR is an internationally widely used framework for describing second language ability via a common 'can do' scale, allowing broad comparisons across languages and qualifications (see Appendix 1 for a figure showing the CEFR global scale descriptors). This purpose relates to some extent to broader issues of recognition of SfL qualifications. For instance, QCA evaluation of the SfL qualifications in 2009 (QCA, 2009) noted a lack of awareness among users and stakeholders of the ways in which the SfL qualifications map to other key known qualifications like GCSEs, EFL qualifications and IELTS as a contributing factor to their lack of external value/currency, the latter 2 being routinely mapped to the CEFR.

We should emphasise that the CEFR mapping carried out in this study can be considered exploratory and preliminary, rather than an 'official' linking, being limited in scope to a subset of qualifications and their components. This was a research exercise, carried out to facilitate our understanding of the extent of comparability of standards between different SfL qualifications, as well as to shed some light on how they compare to other related qualifications such as ESOL International in terms of content and standards at different levels (via the CEFR link).

Why aligning the SfL qualifications to the CEFR may be helpful in understanding their assessment and performance standards

As already noted, at the highest level, both SfL qualifications and ESOL International qualifications have a similar broad purpose of enabling learners whose first language is not English to develop their English language skills. In this way, they are both different from the FSE and GCSE English qualifications, which are intended for learners whose first language is English.

In the context of ESOL International qualifications, the English skills are described in terms of the CEFR levels. The CEFR levels are also commonly used for stating language requirements for different types of visas but are also sometimes used for people who are already in the UK to evidence their level of English, for instance, for entry into higher education. To enter the UK as a skilled worker or a health and care worker, CEFR level B1 is required, while B2 is typically required for entry to higher education institutions. Such uses indicate that there is a common understanding of the broad meaning of the CEFR levels in the context of immigration policy, among some employers and higher education institutions.

DfES (2003) provides a notional mapping between various qualifications and standards frameworks, including NQF, NSAL and the CEFR. According to that mapping, the NSAL levels Entry 1, Entry 2, Entry 3, Level 1 and Level 2 map onto CEFR levels A1, A2, B1, B2 and C1 respectively. While this mapping provides some indication of where the standards for the qualifications such as SfL, developed with reference to the NSAL, should be, this was not an empirical mapping, and did not involve looking at actual assessment and performance standards of individual qualifications developed within those frameworks. Therefore, it might be possible that, for instance, the learners who pass Entry 3 SfL qualifications actually do not exhibit B1 level of performance in terms of the CEFR, or indeed, that they exhibit a higher level of performance than B1 (cf. Curcin & Black (2019) for findings that show a mismatch of RQF and CEFR levels in several GCSE MFL qualifications).

The SfL qualifications are currently not explicitly mapped to the CEFR, but only to the RQF and the NSAL. In addition, as demonstrated in Study 3, the mapping to the NSAL and the core curriculum may not be entirely transparent, appropriate, or comparable between different SfL qualifications. It is certainly not entirely transparent to non-specialists what this mapping substantively means in terms of the skills that the learners at different NSAL/RQF levels have in relation to what might be required for different types of employment or further or higher education, and how those skills relate to the skills of ESOL International learners described in terms of

the CEFR levels, who might be competing for those same jobs or academic courses upon entry to this country.

Given that the broadest purpose of the SfL qualifications is to impart English language skills to learners living in the UK whose first language is not English, it seemed a useful exercise to understand in more concrete terms how individual SfL qualifications align with the CEFR (against which Ofqual regulates ESOL International qualifications) and thus other CEFR-mapped qualifications with similar broad purposes. We wanted to find out if they do indeed impart the B1 level skills at Entry 3 and the B2 level skills at Level 1, for instance. If they did, then English skills at those levels should enable SfL learners to progress to employment or education that requires those skill levels in the same way as English skills at those levels enable international health and care workers (B1) or international students (B2) to access the relevant jobs or education in this country.

Potential tensions and discrepancies in the mapping

Even though the CEFR is intended to provide a metalanguage for description of second language competence across different languages and qualifications, it is not intended to be used indiscriminately and without regard to local context and local educational aims. Any language learning programme or specification of a communicative repertoire for assessment that uses the CEFR would necessarily select aspects of what is described in the CEFR that align with its purpose and the needs of the learners.

It should be noted that the CEFR was not developed with migrant populations in mind. It assumes a level of literacy, such as knowledge of the Latin alphabet, that certain refugees, for instance, may not possess in the target or their native language (ALTE, 2016). Little (2012) gives an example of elaboration of CEFR A1.1 level for migrant learners of French, which included many vocabulary items that are not high frequency (as in the default CEFR A1 descriptor) but are nevertheless essential for adult migrants because they refer to administrative and other procedures that migrants need to be able to cope with.

While the specific context of the SfL qualifications is different to that of ESOL International qualifications and other language qualifications typically mapped onto the CEFR, the fundamental aspects of language competence, which are encompassed in the CEFR and in the core curriculum, are substantively similar, and should be comparable, as confirmed by our subject experts in Study 3. Indeed, the basic communicative approach to teaching and learning language skills in the core curriculum is very much in line with the way language competence, as well as

teaching and learning, are conceptualised in the CEFR.[12] Helpfully, the content of the SfL qualifications themselves is laid out in some detail in the curriculum as well as in individual AO specifications, which aids comparisons. Nevertheless, given the specifics of the SfL context, in our CEFR mapping for the SfL qualifications we acknowledge key points of alignment as well as discrepancies and tensions.

Method overview

The approach in this study followed the methodology used in Curcin & Black (2019) and 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, 2020). It was designed to provide empirical evidence for a link between performance and assessment standards of the 4 SfL qualifications in our sample to the CEFR.

Due to resource limitations, as well as availability of samples of student work, the study was carried out on a subset of qualification levels and components (reading comprehension and writing) across the 4 AOs. Therefore, it is only indicative for levels and components not included in this study.

Following the Manual, the study involved 5 stages:

1. familiarisation/training of participants
2. content mapping (relating the construct/content of the SfL qualifications to the CEFR) at Entry 1 and Entry 3
3. linking of assessment standards for reading comprehension at Entry 3
4. linking of performance standards for writing at Entry 1, Entry 3 and Level 1
5. empirical validation and evaluation of stages 3 and 4 in particular

A panel of 11 participants was recruited for the study. Of these, 10 had extensive CEFR experience, mostly in the context of EFL/ESOL International qualifications, while 3 also had expertise in SfL qualifications in the context of teacher training, curriculum development and Ofsted inspections. The participants included independent language assessment consultants, teachers, examiners and academics affiliated with higher education institutions. They had wide-ranging experience of

[12] The methodological message of the CEFR is that language learning should be directed towards enabling learners to act in real-life situations, expressing themselves and accomplishing tasks of different natures. Thus, the criterion suggested for assessment is communicative ability in real life, in relation to a continuum of ability (Levels A1-C2). This implies that the teaching and learning process is action-oriented, and suggests planning backwards from learners' real-life communicative needs, with consequent alignment between curriculum, teaching and assessment (Council of Europe, 2020).

working on test development or CEFR alignment projects across different language testing institutions in the UK and internationally, as well as teaching and examining in the context of ESOL International qualifications. All except one participant had significant experience of the CEFR in either teaching/examining or test/resource development, and some in both areas. One participant had some experience of it through teaching only. All except 2 participants had some or significant experience of developing reading comprehension tasks targeted at specific CEFR levels and had taken part in similar standard linking exercises before. None of the participants were permanently employed by any of the AOs involved in the study, although some of them worked for some of those AOs as consultants.

Following familiarisation and training (stage 1), 3 participants carried out the content mapping (stage 2) ahead of the standard linking activities. All 11 participants then took part in the linking of assessment standards for reading comprehension (stage 3) and in the linking of performance standards for writing (stage 4).

The SfL materials used in the study were provided to us by the relevant AOs. They included current specification documents, mark schemes and any accompanying marking guidance/criteria where appropriate, 2 live assessment instruments per component, and samples of writing from recent examinations.

The study also involved a range of assessment materials and performances from examinations that are already mapped to the CEFR. These materials were used during familiarisation as well as in the standard linking of writing components. We refer to these materials henceforth as the CEFR benchmarks. The CEFR benchmark reading and oral comprehension tasks, were obtained from the Council of Europe website. The CEFR benchmark writing performances were obtained from several sources:

- the Council of Europe website
- Cambridge Assessment teacher handbooks for A2 (Key), B1 (Preliminary) and B2 (First) exams (Cambridge Assessment English, 2016a, b; 2019a, b, c)
- CEFTrain project
- The British Council Aptis test
- ESB ESOL International tests at B1 and B2
- Trinity College London ESOL International tests at A2, B1 and B2

A small set of benchmarks from ESOL Exemplars (DfES, 2004) was also used in the mapping of the writing components.

We present more detailed methodology and results for each of stages 2, 3 and 4 of the study in turn below. Before that, we present some details about stage 1 of the study, which involved familiarisation and training of the participants.

Stage 1: Familiarisation and training

Prior to undertaking any of the linking 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 relevant SfL assessments, as well as the standard linking methodology to be used. This aimed to ensure the integrity and quality of panellists' judgements.

Most familiarisation activities were 'self-study', conducted online in participants' own time over a designated period, using a survey tool set up with a range of activities. All the participants completed familiarisation activities related to global CEFR scales, reading comprehension and writing (one and a half days self-study, ahead of the content mapping or standard linking as appropriate). The participants who took part in content mapping also completed familiarisation activities in relation to speaking and oral comprehension, as well as more detailed familiarisation with SfL curriculum and specifications (2 days self-study).

Individual self-study familiarisation included the following key aspects for all the participants:

- reading of excerpts from the CEFR (2001) which briefly described what the CEFR is, its conceptualisation of language ability, what illustrative descriptors are and how to read them, the meaning of threshold and 'plus-levels' and nature of pre-A1 second language ability
- familiarisation with the global CEFR scale, sorting individual CEFR descriptors from the CEFR global scale into levels, and self-assessment of participants' own CEFR level in a foreign language using CEFR descriptors
- familiarisation with the CEFR scales related to reading comprehension and written production, interaction and mediation
- familiarisation with the concepts of task and question demand vs. empirical 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 demand or difficulty, and consideration of which aspects of stimulus texts and test questions could be manipulated to change their level of demand
- consideration of the most salient aspects that distinguish between reading comprehension tasks targeted at different CEFR levels, using exemplar tasks with known CEFR levels
- familiarisation with SfL reading comprehension assessment materials, including sketching answers to each question paper

- exercises in rating reading comprehension CEFR benchmark tasks and SfL reading comprehension tasks using the Basket method (see stage 3 section below for more details on this method)
- consideration of examples of written performances with known CEFR levels and key features that distinguish between performances at different CEFR levels
- rating written performances with known CEFR levels
- familiarisation with SfL writing assessment materials
- exercises in ranking written CEFR benchmark and SfL performances using the online Cambridge Assessment CJ platform (see stage 4 section below for more details on the rank ordering method)

The participants who took part in the content mapping ahead of standard linking also did the following activities:

- familiarisation with oral comprehension and spoken production and interaction CEFR scales
- consideration of examples of spoken performances with known CEFR levels and key features that distinguish between performances at different CEFR levels
- rating speaking performances with known CEFR levels
- consideration of the most salient aspects that distinguish between oral comprehension tasks targeted at different CEFR levels, using exemplar tasks with known CEFR levels
- exercises in rating oral comprehension CEFR benchmark tasks
- familiarisation with SfL specifications for speaking and listening and with the core curriculum

Following the self-study activities, ahead of standard linking for the reading comprehension components, a half-day discussion meeting was conducted online using MS Teams, where feedback on the activities related to standard linking for reading comprehension was provided and participants were able to discuss their responses. The meeting was attended by all the participants and it was recorded for quality assurance purposes. The discussion meeting included:

- clarification of any concepts and issues from individual familiarisation activities
- more detailed description of the standard linking procedure
- discussion of the ratings collated from the survey tool, 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

There was no separate face-to-face meeting ahead of the standard linking for writing because this was carried out using CJ methodology, which does not normally require face-to-face standardisation (cf. Black & Bramley, 2008; Curcin, Howard, Sully & Black, 2019; Jones, 2009). The CJ methodology does not require intimate knowledge of a rating scale and, by the very nature of the method, helps to cancel out severity or leniency effects in judgements, which is why it was chosen for use in this study.

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 participants' responses to our familiarisation evaluation questions suggest that they generally considered the activities effective in helping them use the CEFR and the relevant linking methodology confidently in the context of the current study.

Stage 2: Content mapping to the CEFR

This exercise involved relating the content, construct and intended demand of Entry 1 and Entry 3 SfL qualifications from the 4 AOs in our sample to the CEFR. It provides the initial, tentative link to the CEFR, based on profiling the content and construct of the qualifications in terms of the CEFR categories and levels. It also, therefore, provides an initial indication of the comparability between different qualifications in terms of their content and construct. The content mapping indicates the CEFR level that could be evidenced by learners given the content and nature of the tasks in a qualification.

As already mentioned, the CEFR assumes and advocates a communicative, action-oriented approach to language learning, and this approach is reflected in the aspects of language competence that are described through its various descriptor scales. As we said, based on the findings of our Study 3, it was deemed broadly appropriate to consider the SfL qualifications in relation to the CEFR, given that the approach suggested in the curriculum for these qualifications is intended to be communicative and action-oriented too. However, the current content mapping study helped to understand in more detail the extent to which it is appropriate to link the assessment and performance standards of the SfL qualifications to the CEFR based on alignment of the 2 constructs. This is particularly important given that the SfL qualifications were not developed explicitly with reference to the CEFR.

Method

Content mapping was carried out by three experts, with each qualification reviewed by 2 experts independently of each other. The experts were asked to review the relevant specification documents and 2 samples of live assessment materials (including exam papers/tasks, mark schemes, audio materials for oral comprehension) per level (Entry 1 and Entry 3) and component (reading, writing, speaking and listening). They were provided with the relevant CEFR reference scales, tables and excerpts from the CEFR text to ensure thorough understanding of the relevant aspects to consider in content mapping. To capture their views in relation to the alignment of the relevant aspects of the content and construct of each qualification to the CEFR, they completed a series of forms adapted from the Manual (ibid.):

- A1 – general examination description
- A9 – oral comprehension
- A10 – reading comprehension
- A11 – spoken interaction
- A12 – written interaction
- A13 – spoken production
- A14 – written production
- A15 – integrated skills combinations
- A17 and A18 – spoken and written mediation
- A19 – communicative language competence in reception
- A20-A22 – communicative competence in interaction and production
- A24 – confirmed estimation by skill/component

The experts were also asked to:

- highlight the most relevant CEFR descriptive scales that apply to each of the components they mapped
- provide an overall estimate of the CEFR level for each SfL component and their justification for it (for instance, summarising across written production, interaction and communicative competence in interaction/production to give their overall rating for an SfL writing component)
- note any general comments about the assessments they reviewed, including any concerns about specific tasks or approaches

- note any significant discrepancies between the CEFR level they assigned and what the CEFR at that level describes in general and consider whether there are any significant obstacles in terms of content, purpose, or any other reasons why an attempt to link the SfL qualifications to the CEFR might be inappropriate

Their completed forms were reviewed by an Ofqual researcher and then discussion meetings were held remotely via MS Teams with the experts to come to an agreement between each pair of experts on any discrepancies in individual views. There were separate discussions for each qualification. The final CEFR ratings were confirmed at the end of each discussion.

Results

The tables below present the results of the content mapping for each of the 4 AOs in our sample, for each of the components at Entry 1 and Entry 3[13]. In some cases, different tasks, task types or a significant proportion of questions within a single component were deemed to be at noticeably different levels (or sub-levels), and these are represented by slashes in these tables. Each of these tables is followed by a table listing the CEFR activities and strategies scales that are relevant for describing each component across most AOs at each level. In the cases where the relevant scales do not apply to all 4 AOs, the number of AOs to which they do apply is noted in brackets.

General observations

As can be seen from the tables below, it was possible to map a range of aspects of language competence assessed in the SfL qualification to the CEFR. This is an important finding as it highlights the essential correspondence between the core curriculum, SfL specifications and the CEFR across the majority of the content.

The specifications and assessments were thought to be mostly communicative and functional in approach and were mostly deemed to involve useful and meaningful tasks, particularly in speaking but also in writing components. In most cases, the amount of language production/interaction required in the speaking and writing components was deemed appropriate for the relevant CEFR level implied. The reading comprehension assessments were deemed to be least in line with a communicative, functional approach.

Across all specifications and components reviewed, all 4 principal domains of language use were represented (namely, public, personal, occupational and

[13] More detailed tables with mappings for each AO and level are presented in Appendix 3.

educational). As can be seen from the tables below, there was a reasonable degree of correspondence between the AOs in relation to aspects of language competence assessed by each. However, some discrepancies between the AOs and components were observed in the nature of the tasks or amount and nature of language required, often consistent with the findings from our studies 1 and 3. These discrepancies are generally reflected in the CEFR level or sub-level assigned to each AO and component and can also be observed from the differences in applicability of different CEFR scales to different AOs and components.

All aspects of the CEFR communicative language competence scales were found to be relevant across all SfL components at both levels. In general, the experts noted that there was little in the way of integrated skills assessed explicitly. Therefore, these are not presented separately in the tables below. There was also very little assessment of mediation, apart from some limited aspects present in the speaking and listening component with some AOs.

As can be seen from the tables below, at Entry 1, most AOs and components map to A1 or A1+. However, speaking at Entry 1 seems to be pitched at a higher CEFR level (A1+/A2) than the other components (A1/A1+), and this is the case across the 4 AOs. This pattern for speaking seems to be in adherence with the level implied in the curriculum according to our Study 3. At Entry 3, most of the AOs and components map onto B1 or B1+ CEFR level.

Table 4 Content mapping at component level – Entry 1

AO	Reading comprehension	Writing	Speaking and listening
AO1	A1+	A1/A1+	L: A2 S: A2
AO2	A1/A1+	A1	L: A2 S: A2
AO3	A1+	A1/A2	L: A1+/A2 low S: A2+
AO4	A1/A1+	A1	L: A1+ S: A2

Table 5 The CEFR scales represented in SfL components – Entry 1

Reading comprehension	**Writing**	**Speaking and listening**
Reading correspondence	**Written interaction:**	**Oral comprehension:**
Reading for orientation	Correspondence	(3 AOs)
Reading for information and argument	Notes, messages and forms	Understanding conversation between other people
Identifying cues and inferring (more clearly represented in 2 of 4 AOs)	**Written production:**	Understanding audio media and recordings
	Creative writing (3 AOs)	Understanding announcements and instructions (1 AO)
		Oral interaction:
		Asking for clarification
		Understanding an interlocutor
		Informal discussion (with friends) (3 AOs)
		Conversation
		Information exchange
		Oral production:
		Sustained monologue - describing experience/ putting a case (2 AOs)
		Mediating a text:
		Relaying specific information in writing (1 AO)

Table 6 Content mapping component levels – Entry 3

AO	**Reading comprehension**	**Writing**	**Speaking and listening**
AO1	B1	B1/B1+	L: B1 S: B1+
AO2	B1+	B1+	L: B1+ S: B1+/B2
AO3	B1+	A2/B1+	L: B1

			S: B1+
AO4	B1/B1+	B1	L: B1 S: B1

Table 7 The CEFR scales represented in SfL components – Entry 3

Reading comprehension	**Writing**	**Speaking and listening**
Reading correspondence	**Written interaction:**	**Oral comprehension:**
Reading for orientation	Correspondence	(3 AOs)
Reading for information and argument	Notes, messages and forms	Understanding conversation between other people
Reading instructions	Goal-oriented online transactions and collaboration (3 AOs)	Understanding audio media and recordings
Identifying cues and inferring (3 AOs)		Announcements and instructions (1 AO)
Explaining data (1 AO)	**Written production:**	**Oral interaction:**
	Creative writing	Turn taking
	Reports and Essays (3 AOs)	Co-operating (2 AOs)
		Goal-oriented co-operation
		Asking for clarification
		Informal discussion (with friends) (3 AOs)
		Obtaining goods and services
		Information exchange (3 AOs)
		Understanding an interlocutor
		Oral production:
		Sustained monologue – describing experience/putting a case/giving information
		Addressing audiences (2 AOs)
		Oral mediation:
		(2 AOs)

	Facilitating collaborative interaction with peers
	Collaborating to construct meaning

Some discrepancies between the AOs

At Entry 1, main differences between the AOs are apparent in writing, where 2 AOs appear to assess at a somewhat higher level in some of their tasks (A1+/A2) compared to the rest (A1).

At Entry 3, some discrepancies were observed across different components. In reading comprehension, AO1 was deemed to assess at a slightly lower level than the others (B1 vs. B1+). AO4's writing assessments appear to be somewhat lower than the others at B1, compared to up to B1+ with the others. In speaking, AO4 appears to assess at somewhat lower level than the others (B1), while AO2 appears to assess at a noticeably higher level (B1+/B2). In addition, one AO does not assess oral comprehension in its speaking and listening components at either Entry 1 or Entry 3 levels, which necessarily affects the nature of language competence evidenced in these assessments compared to the other 3 AOs in the sample.

Discrepancies with the CEFR

Where the specifications and assessments departed from the CEFR this was mostly in terms of 3 key aspects:

1. occasional use of assessment approaches and task types that do not elicit communicative language competence sufficiently well
2. approach to literacy assessment in relation to:
    - the assessment of Latin alphabet knowledge, which is assumed rather than described in the CEFR at the corresponding levels
    - the requirement of a degree of accuracy in spelling and punctuation, and an awareness of the writing process, not expected at corresponding CEFR levels
3. approach to assessment of oral comprehension

The first category of discrepancies involves, for instance:

- limited use of sufficiently interactive writing and speaking tasks in some assessments
- occasional assessment of knowledge of grammar or spelling through non-functional non-communicative gap-filling or selected response tasks

within the writing component, which may be more appropriate for instructional use

- assessment of alphabetical ordering in a context-free way rather than in relation to application in using reference materials
- assessment of declarative knowledge of text features (comprehension of words such as: bullet points, diagrams, columns, titles, sub-headings) or grammatical features, rather than assessment of their functional use or how they contribute to text meaning
- assessment within a single topic area across all components at the same level

These discrepancies represent a departure from what might be considered best practice in terms of action-oriented communicative approach typically assumed by the CEFR. While these discrepancies do not entirely preclude the possibility of mapping to the CEFR, they do limit the interpretation of the indicative levels in terms of how much evidence they elicit of learners' communicative ability in relation to some of the above-mentioned aspects. It should be noted, however, that such issues were only present in the minority of AOs reviewed, and usually only in a subset of their components (mostly in reading comprehension). Most of those issues were noted by the subject experts in our study 3 as well, as aspects that were at odds with the core curriculum.

The mismatches in categories 2 and 3 make the CEFR content mapping fuzzier, reflecting the specific context of the SfL qualifications. The discrepancies in the second category most notably include assessment of knowledge of Latin alphabet in Entry 1 reading comprehension components. This particular aspect cannot be mapped to the CEFR as the descriptive scales assume rather than elaborate on this knowledge. Furthermore, the fact that this is explicitly assessed through a subset of questions within the Entry 1 SfL reading comprehension tests, while at the same time assuming that the learners can engage with and read the other reading comprehension questions as well as the stimulus texts, makes it unclear why such explicit alphabet-related questions are at all necessary within the assessments at this level. Arguably, they may be more appropriate at pre-entry levels or only for instructional/formative purposes.

Furthermore, some Entry 1 assessments, which are broadly at A1/A1+ CEFR level, appear to require a degree of orthographic accuracy (such as spelling, use of capitalization) that do not appear until B1 level in the CEFR. In addition, in the CEFR, the process of writing is not addressed explicitly, whereas in Entry 3 SfL specifications for writing, the following are very clearly specified and required within some assessment criteria:

- the process of planning and drafting when writing

- awareness that some types of writing do not need formal planning while others such as important letters or college assignments do
- proofreading and correcting writing for grammar and spelling

While the CEFR does include reference to some of the above-mentioned aspects in relation to pragmatic and sociolinguistic appropriateness, as well as in its orthographic control scale, or under coherence and cohesion, these are not always explicitly related to writing only, and are couched in terms of achieving communicative aims rather than being the focus of assessment.

In the context of the SfL qualifications, such requirements in the domain of literacy are related to their link to the NSAL, which are primarily targeted at learners for whom English is the first language. The results of our Study 3, as well as this CEFR content mapping, suggest that perhaps such requirements may not be entirely appropriate at all language ability levels in the context of learning English as a second language, and, in particular in the migration context. There are suggestions in the literature (ALTE, 2016) that in the migrant context, the focus on linguistic accuracy may not be as relevant as the ability to get the main idea across. Arguably, passing a test should be determined by how much the candidates have understood the task and how appropriately they completed it with more attention placed on pragmatic aspects, and less on grammatical ones. This may be particularly true of learners up to and including B1 level where language learning tends to be for general communicative purposes and provides a basis for the development of more advanced proficiency (B2-C2) that increasingly interacts with academic, vocational and professional use of the target language (Little, 2012).

The CEFR, and the ESOL core curriculum to an extent, appear to recognise this more clearly than the NSAL. This potentially raises questions about whether mapping English as a second language qualifications, such as SfL, to standards intended for English as the first language qualifications is entirely appropriate.

The content mapping also demonstrates that there is not a great deal of emphasis on oral comprehension in the SfL specifications and assessments, which we highlight as the third category of discrepancies. Oral comprehension is defined in the CEFR as one-way comprehension, for instance as a member of an audience or comprehension of audio media and recordings. The SfL specifications and assessments place more emphasis on assessing oral interaction. This, in itself, may not necessarily be seen as a limitation of the SfL specifications and assessments, as long as it is consistent with the agreed purposes and the curriculum. It was noted in the results of Study 3 that these approaches are not entirely aligned with the core curriculum either, as oral comprehension gets a reasonable amount of attention in the curriculum.

Implications of assessment quality issues for the CEFR mapping

In this content mapping exercise, the experts were asked to consider any particularly prominent issues in task or question design that might limit the interpretation of the mapping. Most issues that were noted relate to:

- occasional tasks not being optimally designed to elicit communicative language competence
- sometimes insufficient prompts in writing and speaking assessments
- sometimes guessable or otherwise poorly constructed reading comprehension questions
- sometimes insufficient comparability of topic demands between assessment versions within an AO and component

Such issues can also limit the interpretation of the CEFR linking, where assessment operationalisation rather than the scope of the content specification or intended demand, constrains the nature of the skills elicited by the assessments thus reducing overall validity of both the CEFR link and the assessment itself.

Furthermore, even though all of the AOs reviewed do have certain quality assurance processes in place to evaluate their assessment instruments and results, these may not always be in line with some of the best practice in assessment development (cf. AERA/APA/NCME, 2014; ALTE, 2016; 2020), for instance, absence of field testing of questions or tasks, or occasional failure to evaluate item functioning in tests. Furthermore, as already discussed, the nature of QA challenges and threats to validity are different in the context of internal vs. external assessment models. This further speaks to the need to improve some of the aspects of the SfL qualifications to ensure their optimal functioning in relation to their purposes and therefore the validity of their results interpretations as well as of the CEFR linking.

Stage 3: Standard linking of reading comprehension components

As already noted, a content mapping exercise can indicate the CEFR level that could potentially be evidenced by the candidates (based on the content, contexts, nature of tasks and assessment criteria), not the pass grade standard. For instance, the indicative CEFR level based on content mapping could in theory be higher than the pass grade standard would be. To establish the CEFR level implied by the standard at the pass grade, standard linking exercises are required.

To get an indication of how SfL reading comprehension components relate to the CEFR in terms of their grading standards, as well as how the (CEFR-linked) pass grade standards compare between different AOs, we linked 4 Entry 3 reading comprehension tests, one from each AO, to the CEFR. 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, which may not support holistic judgements of performance quality. Following Brunfaut & Harding (2013), we used a modified version of a standard-setting procedure called the Basket method to conduct standard linking.

Method

The Basket method

The Basket method requires the participants in a standard-linking panel to rate each task[14] in a test 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 into a CEFR level 'basket'. The Modified Basket method additionally requires panellists to select a CEFR sub-level for a task after deciding on the CEFR level first. The sub-levels are based on the CEFR scales, with low, mid and high sub-categories within each level.

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

Level	Sub-level
Above B2	Above B2
B2	B2 high
B2	B2 mid
B2	B2 low
B1	B1 high
B1	B1 mid
B1	B1 low

[14] In reading comprehension tests, it is usually necessary to consider the demand of a specific question in conjunction with the relevant parts of the stimulus text which the question refers to, any intentionally distracting parts of the text, and the interaction between these elements. Here, and throughout, we refer to this as 'task'. 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 achieve each possible creditworthy mark point.

A2	A2 high
A2	A2 mid
A2	A2 low
A1	A1 high
A1	A1 mid
A1	A1 low
Pre-A1	Pre-A1

In order to rate each task in relation to the CEFR, each participant needed to answer the following key question:

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

During their familiarisation, it was explained to the participants that, in general, global CEFR level descriptors summarise the threshold (criterion) abilities of learners that belong to a level (North & 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 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.

To answer the standard linking question for each task or mark point, each participant 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 to allow them to answer correctly or not. It was highlighted that this judgement did not imply that learners at a lower level could not give the correct response. It meant 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. Therefore, 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 (how this task may have performed on the actual test). Furthermore, statistical information about question difficulty was not available to us to consider alongside the panel ratings.

Procedure

All 11 participants took part in the standard linking after familiarisation. They reviewed and rated 4 live Entry 3 reading comprehension tests, provided to us by the AOs.

The standard linking activities were conducted online, in 2 rounds. First, the participants were given about a week to complete their independent ratings for each of the 4 tests and submit the ratings via an online survey tool. These constituted the initial round of ratings.

The initial ratings were then collated and the distribution charts of ratings for each task and mark point created to present to the participants for discussion. An online discussion panel meeting was then held over 2 half-days during which the initial ratings for each test in turn were discussed. The discussion was chaired and facilitated by an Ofqual researcher. During the discussion, the panellists were asked to justify their ratings, and to consider the rationale of other panellists for theirs. The panel sessions were video recorded with the consent of the participants.

After the discussion, the panellists were asked to consider their ratings again independently and to amend any ratings where they 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 imperative to agree with the others. This second round of ratings then constituted the final round. These ratings were used as the basis for calculation of the pass cut score on each test which would represent the relevant CEFR level, as described below.

Data analysis

In order to calculate the mean ratings across panellists (henceforth, raters) for each task and mark point, following Brunfaut and Harding (ibid.), the sub-level scale with low, mid and high categories was transformed into a continuous numerical scale shown in Table 9.

Table 9 Numerical rating scale categories – CEFR sub-levels

Sub-level	Numerical score
Above B2	4.67

B2 high	4.33
B2 mid	4
B2 low	3.67
B1 high	3.33
B1 mid	3
B1 low	2.67
A2 high	2.33
A2 mid	2
A2 low	1.67
A1 high	1.33
A1 mid	1
A1 low	0.67
Pre-A1	0.33

Individual rater scores were transformed into the numerical scores from Table 9. These were then averaged across raters to get a mean score for each task and mark point. The mean scores were then transformed back into the CEFR sub-level as well as the CEFR level ratings, using the ranges shown in Tables 10 and 11. 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 midpoint of each of these parts as the scale point. A similar logic was applied to the sub-levels.

Table 10 Task score ranges - CEFR sub-levels

Sub-level	**Score range**
Above B2	>4.51
B2 high	4.17-4.50
B2 mid	3.84-4.16
B2 low	3.51-3.83
B1 high	3.17-3.50
B1 mid	2.84-3.16
B1 low	2.51-2.83
A2 high	2.17-2.50

A2 mid	1.84-2.16
A2 low	1.51-1.83
A1 high	1.17-1.50
A1 mid	0.84- 1.16
A1 low	0.51-0.83
Pre-A1	0-0.5

Table 11 Task score ranges - CEFR levels

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

The test cut scores were calculated based on sums of weighted frequencies of the mean CEFR level ratings. The approach taken for calculating the cut scores followed Brunfaut & Harding (ibid.) and de Jong (2009) in using weightings that imply a 'comfortable' achievement of a particular level. 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 (Brunfaut and Harding, ibid.; de Jong, ibid.). This approach was taken 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).[15]

For instance, based on the following frequencies of tasks rated at each CEFR level (Table 12), 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 12 Example frequency table

CEFR level	N marks	Cut Score

[15] 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 example from table 10, this would mean that A2 cut score would be 2, and B1 cut score 38.

A1	1	1
A2	37	20
B1	22	42

The mean task sub-level ratings were used to categorise the tasks within each test in relation to their content specification and to consider whether the demand implied by the CEFR sub-level of each task was sufficiently well targeted to ensure that demand distribution of the tasks was appropriate across those targeting specific ACs, as well as across the whole test. It was important to consider CEFR task demand targeting at AC level as well as at the whole-test level because, as described in the Assessment review section, some of the reading comprehension tests apply the mastery model, whereby the learners have to get a certain number of marks for each AC in order to get the pass grade rather than just a particular compensatory test-level score.

Results

Evaluation of judgement consistency

The analysis of the results of the standard linking for the reading comprehension tests involved checking that the outcomes are based on reasonably reliable and consistent rater judgements. This was established by calculating intra-class correlations (ICCs) (Hallgren, 2012) on individual rater scores.

We report both agreement (showing agreement level in absolute terms) and consistency (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 the initial ratings. It can be seen from Table 13 that even based on the initial ratings the ICCs were reasonable, and in some cases very high. This suggests that our participants were able to use the CEFR scales independently with a high level of agreement and consistency.

Table 13 ICCs based on initial ratings

Test	ICC agreement	Confidence interval	ICC consistency	Confidence interval
AO1	0.56	0.22 < ICC < 0.80	0.62	0.29 < ICC < 0.84

AO2	0.61	0.34 < ICC < 0.80	0.73	0.53 < ICC < 0.87
AO3	0.73	0.50 < ICC < 0.89	0.77	0.56 < ICC < 0.91
AO4	0.95	0.92 < ICC < 0.97	0.96	0.93 < ICC < 0.98

We also calculated the ICCs for the final ratings, given the panel discussions – presented in the table below. It can be seen in Table 14 that agreement and consistency levels of the ratings improved after discussion. The cut score and linking analysis is based on these final ratings.

Table 14 ICCs based on final ratings

Test	ICC agreement	Confidence interval	ICC consistency	Confidence interval
AO1	0.82	0.67 < ICC < 0.92	0.84	0.70 < ICC < 0.93
AO2	0.89	0.81 < ICC < 0.95	0.92	0.86 < ICC < 0.96
AO3	0.90	0.82 < ICC < 0.96	0.91	0.82 < ICC < 0.96
AO4	0.98	0.97 < ICC < 0.99	0.98	0.97 < ICC < 0.99

Linking to the CEFR

In this section, we present the outcomes of the linking exercise in relation to percentage scores and proportions of tasks/marks in each test belonging to different CEFR level or sub-level categories. Figure 1 shows the distribution of task sub-level ratings, while Table 15 shows the mean CEFR sub-level rating, standard deviation and range for each test. It can be observed from Figure 1 that most of the tasks across the 4 tests are rated B1 low to B1 mid (cf. Table 11). However, in some cases there is more of a spread across the A2 and B1 levels, and in one case there are also tasks at A1 level alongside some B1 high tasks.

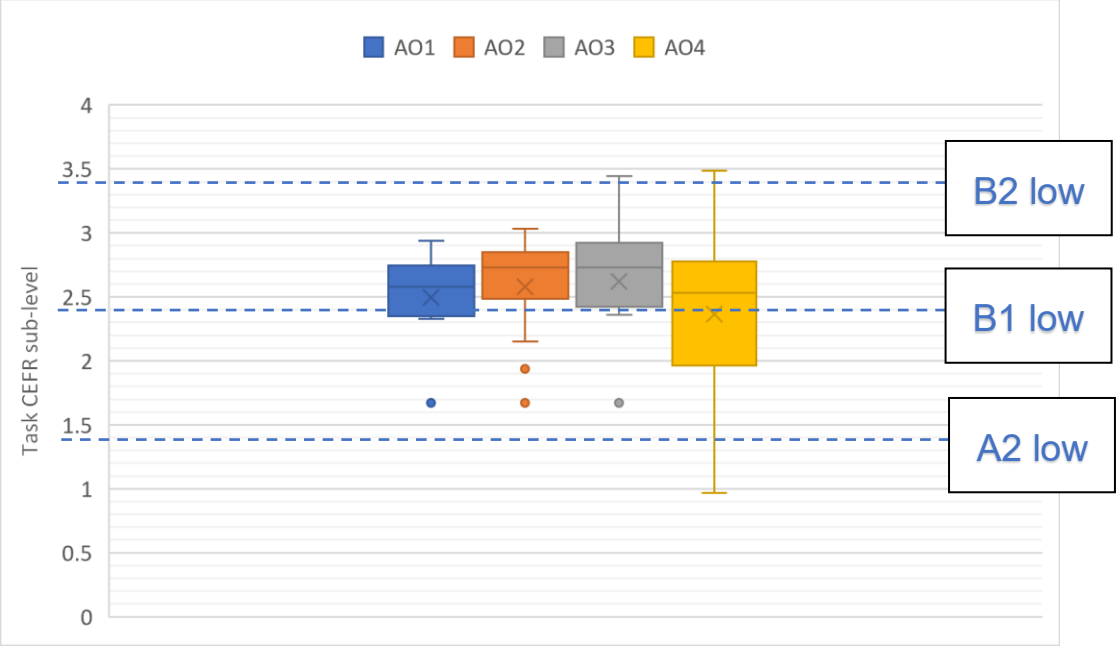


Figure 1 Box plots[16] showing task sub-level rating distribution by test

It can be seen from Table 15 that the AO2 and AO3 tests have the means within the B1 score range, while the AO1 and AO4 tests have the means at the top of A2 range, suggesting that these tests are somewhat less demanding overall.

Table 15 Descriptive statistics for task sub-level ratings by test

Test	Mean CEFR rating	SD	Range
AO1	2.50	0.37	1.62-2.94
AO2	2.58	0.39	1.67-3.03
AO3	2.63	0.43	1.76-3.46
AO4	2.38	0.68	0.97-3.48

The indicative CEFR mapping is shown in Table 16 below. Column 2 shows the CEFR (compensatory) cut scores as the percentage of total marks on each test that would be required to demonstrate B1 level of performance. The CEFR cut scores were calculated based on sums of weighted frequencies of the mean task CEFR

[16] In these box plots (aka box and whisker plots), the box shows the middle 50% of mean task sub-level ratings (the interquartile range) and the whiskers represent the ratings outside of the interquartile range, extending 1.5 times the interquartile range above and below. The larger the box and whiskers, the greater the variability in the task ratings within the test. Tasks with mean ratings that fall outside of the whiskers are shown as solid points and are considered to be outliers. The solid line within each box shows the median value – the midpoint of the task mean ratings within each test (in other words, 50% of the tasks in the test fall above and 50% of tasks fall below this rating value).

level ratings, as described in the Data analysis section[17]. These cut scores, alongside the task demand profile shown in Figure 1 above in terms of mean sub-level CEFR ratings, suggest that the 4 tests are broadly comparable in terms of their indicative CEFR level. However, these CEFR cut scores are derived in a compensatory way and cannot be directly interpreted as indicative of the CEFR level implied by the actual operational cut scores of these tests.

This is because the actual operational cut scores are not all compensatory and differ in important ways in how they are defined. Given different approaches to defining the pass grade cut scores, with some AOs having AC-level, section-level or item-level hurdles, sometimes with additional compensatory test-level cut scores and sometimes without, it is necessary to consider the demand profile of items targeting individual hurdles and what this reveals about the indicative operational CEFR level.

To enable an approximate comparison of the indicative CEFR cut scores and the operational cut scores/hurdles, we cross-referenced the individual tasks with the relevant test specifications, where each task is typically assigned to a specific AC, test section or other hurdle where these are used. For each test, we then calculated the proportion of tasks/marks at each CEFR sub-level which would need to be answered correctly to pass each hurdle and achieve the operational cut score (however this cut score was defined). We assume that learners would be more likely to answer any items at a lower CEFR level correctly and less likely to answer higher-level items (consistent with the weightings for calculating the CEFR cut scores). Recall that the method for calculation of the CEFR cut scores assumes that 50% of the B1 tasks contribute to the cut score in addition to any lower-level tasks (column 3 in Table 16).

Column 4 in Table 16 shows which proportion of the total number of B1 level tasks/marks in each test needs to be answered correctly to pass the relevant hurdles and achieve the operational cut-scores (the remaining tasks that need to be achieved are A2 or lower). It can be seen that, except for the AO1 test, over 50% of the B1 tasks in each test need to be achieved to achieve the cut score and/or different hurdles. However, in the AO1 test, only 20% of the B1 tasks in the test need to be answered correctly to achieve the hurdles.

Table 16 Indicative CEFR mapping

[17] Given the profile of task CEFR ratings shown in Figure 1, cut scores for A2 and, in one case, A1 levels could be set based on this data. However, these are not substantive cut scores given that the number of tasks targeted at these levels is not sufficient to evidence A2 level of competence. These tests are intended as single-level tests, with only pass and fail grades at each level, and are targeted at B1 level according to the content mapping.

Test	Indicative CEFR B1 cut score (compensatory)	Proportion of tasks at B1 level contributing to the CEFR cut score	Proportion of tasks at B1 level contributing to the operational cut score[18]	Indicative operational CEFR level
AO1	62%	50%	20%	A2
AO2	59%	50%	55%	B1
AO3	62%	50%	53%	B1
AO4	68%	50%	53%	B1

The indicative operational CEFR levels for each test, based on taking into account the proportion of B1 tasks contributing to the operational cut scores, are shown in column 5. Considering the CEFR demand profile of the tasks required to achieve operational cut scores/hurdles, it can be concluded that the operational criteria for the pass grade translate into B1 CEFR level for 3 of the 4 tests, with the proportion of B1 items likely to contribute to achieving the relevant hurdles similar to the proportion contributing to the corresponding compensatory CEFR cut scores. Indeed, the CEFR-linked percentage cut score is lower than the percentage cut score at operational pass grade boundaries for these 3 tests (the operational percentage cut scores are not presented here for anonymity reasons). With the AO1 test, the minimum pass criteria do not translate into reliable evidence of B1 level, given that only a very small proportion of items at B1 level may be required to achieve the pass grade. This is despite the fact that the overall profile of this test and the implied compensatory CEFR cut score do align with B1 level, similarly to the other 3 tests. If the operational criteria for achieving the pass grade on the AO1 test were brought in line with the other tests, the standards on all 4 tests would be broadly comparable and at B1 level.

Some caveats

There are some caveats to these conclusions. It was noted in the content mapping results section that some of the assessment approaches in these tests may not elicit evidence of the construct of reading comprehension in line with the CEFR and its action-oriented communicative approach. This particularly relates to the following 2 aspects:

[18] The proportion of B1 level tasks is calculated across all hurdles rather than for individual hurdles.

- assessment of alphabetical ordering in a context-free way rather than in relation to application in using reference materials
- assessment of declarative knowledge of text features (words such as: bullet points, diagrams, columns, titles, sub-headings) or grammatical features, rather than assessment of their functional use in terms of how they contribute to text meaning

The former aspect is typically specified in terms of ACs that themselves suggest context-free non-functional approach to assessment (for instance, “Use first and second placed letters to order words in alphabetical order”). These tasks were rated as A2 low in all tests, though this mapping is not entirely appropriate and is there to indicate potential demand of such tasks were they to use an approach more in line with the CEFR, in relation to the “Reading for orientation” scale and its low A2 descriptor “Can locate specific information in lists and isolate the information required (for instance, use the Yellow Pages to find a service or tradesman)”.

The latter aspect is specified in the ACs that suggest communicative approach (for instance, “Use language features to identify meaning in short straightforward text”), in line with the CEFR and its “Identifying cues and inferring” scale. Therefore, in the latter case, the misalignment appears to stem from assessment design choices rather than construct definition. Furthermore, the requirements for declarative knowledge sometimes raise the demand level of these tasks beyond what might be considered appropriate for B1 level, though not in relation to reading comprehension construct, but rather in relation to knowledge of linguistic vocabulary denoting text and grammar features.

While these discrepancies do not entirely invalidate the mapping to the CEFR, they do limit the interpretation of the indicative levels observed in terms of how much evidence they entail of learners’ communicative ability in relation to the above-mentioned aspects. Furthermore, these assessment approaches were noted by the subject experts in our study 3 as being at odds with the spirit and intention of the core curriculum too. In this sense, these discrepancies represent a departure from what might be considered best practice in terms of the CEFR action-oriented communicative approach as well as in terms of the curriculum. These assessment approaches should therefore be reconsidered to increase the overall validity of the way in which these tests elicit the evidence of the intended constructs. It should be noted that similar assessment approaches are used across all 4 tests, and therefore do not impact on the conclusions about comparability of standards between them.

In addition to these issues, it was already noted in our study 1 that a number of items across all tests sometimes had implausible distractors or guessability and other issues. Some of these issues were also flagged by the subject experts and the CEFR experts. The CEFR mapping additionally demonstrated that some of the items are poorly targeted to the intended demand at the cut score and do not contribute

appropriately to the reliability of the pass grade, which is not intended to differentiate between multiple competence levels. More items closer to the intended demand at the pass grade threshold would help increase test reliability and classification consistency. Where tests are fairly short, this compounds other issues that could potentially reduce test reliability and classification consistency (for instance, guessability). While these are all general assessment design issues that could be improved, they do currently reduce the validity of test score interpretation, whether in relation to the CEFR linking or more generally.

Stage 4: Standard linking of writing components

The fourth stage of the linking study was to relate the pass grade performance standards of the writing components from the 4 SfL AOs to the CEFR and to each other. We used the rank ordering comparative judgement method to do this, as previously advocated by Jones (2009) and implemented in Curcin and Black (2019). The focus of this stage was on Entry 1, Entry 3 and Level 1 writing.

Jones (ibid.) showed that the method can be used to replicate the results of standard setting panel ratings using CEFR descriptor scales. In the current study, the method was used to create a common scale of quality across SfL performances with unknown CEFR levels and selected CEFR benchmark performances where the CEFR levels were already known. In this way it was possible to estimate the CEFR level of the SfL scripts from their position on the common scale relative to the CEFR benchmarks.

Method

Rank ordering

Rank ordering is a comparative judgement (CJ) technique for capturing judgements of quality or other subjective attributes of different kinds of artefacts. Previous research exercises have found that rank ordering, alongside other CJ methods, is a valid method for comparing performance standards in educational assessment, for example, between examinations from different years (Bramley, 2007; Black and Bramley, 2008; Curcin, Howard, Sully and Black, 2019; Benton, Leech and Hughes, 2020; Benton, Cunningham, Hughes & Leech, 2020).

The theory underlying the rank ordering method and other CJ techniques is Thurstone's law of comparative judgement (Thurstone, 1927). In essence, a function of relative frequency of occasions on which one artefact is judged to be better than another is an estimate of their separation on the scale (of quality) to be created by the analysis (cf. Bramley, 2007). In CJ methods effects such as differential severity

of judges cancel out, while collecting judgement data on a large scale maximises the reliability of the quality scale that is created (Verhavert et al., 2019).

When using this method for comparing assessment performance standards, sets of candidate performances (scripts) from 2 or more examinations are rank ordered in terms of overall quality by multiple judges (such as examiners, subject experts). These rankings are then analysed using a latent trait model (for instance, Bradley-Terry or Plackett-Luce) to place each script onto a single scale of quality. The result of a rank ordering exercise is a common script quality scale across different artefacts (in this case, scripts from different examinations) that were included in the rank ordering exercise.

In the current study, scripts from SfL writing examinations across different AOs and levels were included alongside benchmark scripts from a range of other English examinations, with known CEFR levels (see more details on this below). Therefore, it was possible to see how the SfL scripts and CEFR benchmarks were distributed on this quality scale, and thus relate the performance standards of SfL scripts to the CEFR scale.

Given that latent trait models allow 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 in terms of number of comparisons per script to result in reasonably reliable estimates of the script quality scale. Furthermore, the method involves comparing performances with one another, and putting them in a rank order, rather than using a rating scale, where detailed common understanding of the rating scale is required by all the participants. In the context of this study, while making holistic judgements of script quality does require familiarity with the basic notions and approach inherent in the CEFR and its way of describing writing competence, it does not require detailed (shared) knowledge of the CEFR writing scales by the participants.

Scripts

As already noted, a range of CEFR benchmarks from different sources was used in this study. The sample was opportunistic, obtained either from publicly available sources such as the Council of Europe website, or from certain testing organisations (for instance, the British Council). We sought to use several benchmarks per level in order to ensure that we could capture some of the variety of what different CEFR levels might look like in actual writing performances, acknowledging that no single performance will exhibit all different features of a CEFR level (Little, 2012).

We ensured that the benchmarks came from mostly general English exams, rather than any that are specifically academic or related to specific occupations, to help comparisons with the SfL samples, which are mostly demonstrating general English writing skills, albeit in a migration context. Some of the CEFR benchmarks represent

"borderline" performances at particular levels, while some illustrate more secure performances within a level according to the standards of their respective exams. In this exercise, we were not interested in the comparability between the benchmarks from different institutions, and the sources are anonymised in the results section.

The CEFR benchmarked writing performances were obtained from:

- the Council of Europe website
- Cambridge English teacher handbooks for A2 (Key), B1 (Preliminary) and B2 (First) exams (Cambridge Assessment English, 2016a, b; 2019a, b, c)
- CEFTrain project
- The British Council Aptis test
- ESB ESOL International tests at B1 and B2
- Trinity College London ESOL International tests at A2, B1 and B2

The table below summarises the numbers of benchmark samples used for each CEFR level in the study.

Table 17 Number of CEFR benchmarks by level

CEFR Level	N scripts
A1	6
A2	6
B1	10
B2	8

A small set of benchmarks from ESOL Exemplars (DfES, 2004) was also used to illustrate standards of performance at different levels of the NSAL/core curriculum. ESOL Exemplars samples exemplifying secure performances at Entry 1 (N=2), Entry 2 (N=2), Entry 3 (N=3) and Level 1 (N=3) were used. All indications of the CEFR or RQF level were removed from both CEFR benchmarks and ESOL Exemplars samples.

The SfL samples were obtained from the respective AOs. They all came from live examinations. AO, centre and candidate information was removed from the scripts, as well as marks and annotations.

Ideally, for our SfL script sample, we wanted to have threshold (borderline) pass grade performances at each level from each AO. We also required the samples to have gone through the AO's quality assurance/moderation and to have had confirmed grades. However, it was not straightforward to ensure that all of the performances in our sample did indeed represent threshold performances. One reason for this was purely practical, and partly influenced by the extraordinary Covid

situation at the time of data collection for this study. Some AOs did not have enough threshold scripts from recent live administrations that fitted our brief and were not able to access their archives due to Covid restrictions in order to get threshold performances from earlier administrations.

The other reason was related to how the pass grade standard is specified by the AOs. As discussed in Study 1 (Assessment review), some AOs employ traditional marking when evaluating writing performances, and therefore set a numerical cut score for the pass grade at each level. For these AOs, it was possible to request the performances at or near the numerical pass grade cut score. However, some AOs assess by means of descriptive, binary assessment criteria (met/not met) and do not involve numerical scores. For these AOs, it was not straightforward to define what represents a threshold performance. We therefore relied on those AOs to send us the samples that they deemed as representative of a borderline pass. In order to be confident that at least some of the scripts from these AOs did represent borderline passes, we also requested some borderline fail samples from each AO. The intention was to check how the borderline fail scripts in our sample relate to the borderline pass scripts, with an expectation that they should be relatively close in quality if they indeed represented borderline performances.

From each AO and for each of Entry 1, Entry 3 and Level 1, we requested 7 scripts at or near the pass grade borderline standard/cut score, and 3 borderline fail samples. For Entry 2, we requested 3 scripts at or near the pass grade borderline standard cut score. While Entry 2 was not the focus of our exercise, we wanted to include some Entry 2 scripts to ensure stronger linking across the entire quality scale from Entry 1 to Level 1.

The table below shows the number of samples per AO and level included in the rank ordering exercise. Note that we did not always receive the requested number of samples from all AOs[19], while receiving some extra samples that we did not specifically request (for instance, Entry 2 fail grade samples). AO4 in particular provided very few samples across the board, hence any conclusions about their results will necessarily be more tentative compared to the others.

Table 18 SfL script sample by level

Level and grade	AO1	AO2	AO3	AO4	Total
E1 Pass	7	7	6	2	22
E1 Fail	2	3	3	N/A	8

[19] Given the current situation with centre closures and difficulties in exam administration due to Covid epidemic, some AOs struggled to provide enough suitable samples by our stated deadline.

E2 Pass	3	3	3	1	10
E2 Fail	1	N/A	2	1	4
E3 Pass	7	7	8	2	24
E3 Fail	3	3	3	2	11
L1 Pass	6	7	5	1	19
L1 Fail	2	3	3	N/A	8
Total	31	33	33	9	106

Judging design

In total, there were 146 scripts to rank order, across the benchmarks (N=40) and the SfL samples (N=106). These were distributed across the judges into sets of 4, with each judge required to rank order 26 or 27 sets. All 11 participants took part in this exercise.

The judging allocation design was created such that it maximised the number of times each script was seen across judges while minimising the number of times each judge saw a particular script. Literature suggests that over 20 paired comparisons per artefact should be sufficient for a reasonable level of scale reliability (Verhavert et al., 2019). Therefore, the design ensured that each script was seen 8 times across different judges and sets, translating into 24 comparisons per script. Each judge saw most of the scripts once, and a maximum of 2 times. Each judge had a unique combination of scripts.

Traditionally, where the aim of a rank ordering exercise is to compare scripts from 2 different examination sessions, the sets might contain equal numbers of scripts from each session. In our study, because the number of scripts was not equal between the benchmarks and the SfL samples, and because it was equally important to ensure reliable judgements for both of these sets of scripts, we did not constrain the design in terms of script type in each set. Therefore, some sets contained 2 benchmarks and 2 SfL scripts, while some contained only 1 or none of one of the categories.

We constrained the design so that scripts that were more than 1 level apart (either on the CEFR or RQF scale) could not coincide in a set. There were no constraints regarding how many scripts at each level were in any set. Therefore, some sets contained all scripts at the same level, while some contained a combination of, for instance Entry 1 and Entry 2 or B1 and B2.

Table 19 Key features of the judging allocation design

N judges	N sets per judge	N scripts per set	N times script seen	N comparisons per script
11	26/27	4	8	24

Procedure

The participants were contracted for one and a half working days to carry out the rank ordering exercise. The exercise was conducted online, using the Cambridge Assessment Comparative Judgement Tool. This platform allows for electronic pdf script files to be uploaded and allocated to the judges according to a pre-specified judging design.

During their familiarisation activities, which preceded the rank ordering exercise, the participants were given the opportunity to access practice scripts and rank order them using the CJ platform. At the end of their familiarisation, they confirmed that rank ordering 4 electronic files feasible on screen, and that the CJ platform was straightforward to use.

The participants were provided with detailed instructions about how to complete the rank ordering task. They were asked to work independently of each other and not discuss their judgements. An Ofqual researcher and IT support were on hand in case of any issues with accessing the CJ tool.

As already noted, the participants did not know the CEFR levels of the benchmarks, nor did they know the marks or grades of the SfL scripts. They were asked to rank order the scripts in each set from best to worst in terms of overall quality. It was emphasised that they were to consider the important aspects of writing competence according to the CEFR approach and CEFR descriptor scales as the basis for their judgements of quality. They were not provided with any mark schemes for individual exams that the scripts came from.

The judges 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, with different tasks and layouts. Furthermore, some scripts (typically the SfL scripts) were handwritten, while some were typed up (typically the benchmarks, though a few of these were also handwritten). The judges were alerted to this in their instructions and advised to try and not allow this to affect their judgement of script quality.

All SfL scripts and most benchmark scripts included at least 2 writing tasks. However, a small number of benchmark scripts included a single task. In such cases, quantity of writing in isolation should not have been used as a criterion when comparing with scripts based on 2 or more tasks. The judges were instructed to try and extrapolate as much as possible about the writing quality from the benchmark

scripts based on a single task, focusing on key aspects of performance rather than the quantity of writing. We advised that, when considering such benchmark scripts, they might assume that the candidate who wrote a response to that task could write a piece of similar quality for another task or on another topic.

In general, the judges were asked to consider the whole work of each candidate. There were 2 exceptions to this. Firstly, most SfL scripts contained a form-filling task. In some cases, this was a basic form requiring candidates to enter some personal information. Given that this would reveal the identity of the candidates, we had to redact most or all of this information. Therefore, in most cases, this could not contribute to the overall judgment, and the judges were asked to focus on other, more substantive, writing tasks in those scripts. In some cases, the forms contained a section requiring an open response. The judges were asked to take those into account when judging.

Secondly, one awarding organisation's SfL writing papers contained a "use of English" task, with several questions focusing mostly on aspects of grammar or spelling. The judges were instructed to take the aspects of performance revealed in that task into account as far as this helped to form a judgement about a candidate's overall grammatical or spelling competence, with an expectation that their overall judgement would mostly be informed by the communicative writing tasks in those scripts.

Data analysis

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 SfL and CEFR benchmark scripts was estimated using the Bradley-Terry model (Bradley & Terry, 1952)[20] implemented in the Sirt package in R (Robitzsch, 2019).

Each script is positioned on this scale in terms of its estimated measure (of quality), which is related to the probability of it being judged better than another script in a paired comparison. The unit of the script quality scale is known as a 'logit' or 'log-odds unit'. 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

B_i = the measure for script *i*

[20] Another approach is to use a model for rank-ordered data, allowing for the constraints imposed by a ranking (Plackett, 1975) as implemented in PlackettLuce R package (Turner et al., 2020). The Rasch formulation of the Thurstone model is a special case of this more general model. We analysed our data in this way too, and the results are presented in Appendix 4.

B_j = the measure for script *j*

After the initial analysis run, some data cleaning was undertaken to improve the fit of some judges or scripts where appropriate. In order to preserve most of the data, we removed misfitting observations (based on highest standardised residuals for individual paired comparisons) rather than all judgements from a judge or script that showed some misfit. Initially, observations with standardised residuals higher than 4 were removed from analyses, and analysis rerun. After that, in order to further improve the fit of a couple of judges and scripts, a few more observations with high residuals were removed and analysis rerun. In total, 34 observations out of 3506 were removed. In addition to this, 2 outlier benchmark scripts were removed from analysis. These were flagged as outliers in terms of the logit measure compared to the other scripts at the same CEFR level. 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.

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

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

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

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.[21] In this context, SSR means "reproducibility of relative measure location" (cf. Winsteps Manual). 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 (meaning that they are of better quality) than the scripts estimated with lower measures.

In general, the decision of whether the SSR 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

[21] Separation coefficient is the ratio of the person true SD (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).

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. 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, or, 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 (Wright & Linacre, 1994).[22] 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 corelation between the test score scale and the measure scale is a way to establish whether the judges in a CJ exercise perceived 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. In the current study, it was only possible to get score-measure correlations on subsets of scripts for which a score or at least a score band was known.

Once the appropriate script quality measures are obtained, their logit scale can be plotted against the SfL levels scale/CEFR scale to observe their relative position and extrapolate the likely performance standards of the relevant SfL pass grade boundary scripts at each level in relation to the CEFR benchmark performance standards.

Results

Evaluation of scale properties and model fit

In order to have confidence in the results of a rank ordering exercise, it is necessary to evaluate the properties of the script quality scale and the model fit. The scale separation reliability (SSR) and separation were high, suggesting that the observed

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

differences between scripts were not due to measurement error. This is shown in the table below.

It can be seen that the BT model overestimates the reliability to some extent. However, the PL estimates are also high and indicative of a reliable script quality scale.

Table 20 SSR and separation coefficients

Model	SSR	Separation
BT	0.97	6.25
PL	0.91	3.18

As already noted in the data analysis section, we have based all of the analyses reported below on the estimates from the BT model. Appendix 5 presents the analysis based on the PL model for comparison, though there were no fundamental differences to the conclusions based on these analyses compared to those based on the BT model. The main difference is slightly less differentiation (more overlap) between the A2 and B1 script measures compared to the BT model, consistent with lower separation and shorter scale.

The overall model fit can be seen in the table below. Model fit was satisfactory, with less than 2% standardised residuals greater than absolute 2 and less than 1% standardised residuals greater than absolute 3.

Table 21 Overall model fit

StRes criterion	N valid observations	N	%
> abs 2	3472	66	2
> abs 3	3472	14	0.4

Individual judge fit was also satisfactory, with all infit and outfit mean squares lower than 1.5, suggesting that the judges were consistent in their judgements. Script fit statistics were also largely lower than 1.5, consistent with the overall satisfactory model fit. There were several scripts that won or lost all of their comparisons, resulting in imputed measures.[23]

In order to get a sense of how well the rank order of scripts produced in this exercise matched the rank order based on the CEFR or RQF level, we calculated correlations between script measures and level, where levels were turned into numerical categories 1-4. The correlations were 0.89 for the CEFR benchmark scripts and 0.85 for the SfL scripts.

[23] Judge and script fit statistics are presented in Appendix 5.

Where original scores or score bands (lower and higher A1), were known, which was for a subset of CEFR benchmarks and for the scripts from one SfL AO, we also ran score-measure correlations for those, separately for each benchmark group and SfL script group. With respect to the CEFR benchmark groups, while these correlations are based on a very small number of scripts in each group (min 2, max 4), they do suggest that the judges were able to discern script differences largely consistent with the original score or score band rank order.

Table 22 Original mark (band) correlation with script measure – benchmark scripts

Benchmark group	**A1**	**A2**	**B1**	**B2**
1	N/A	N/A	0.95 (N=3)	0.98 (N=3)
2	N/A	N/A	1.00 (N=2)	-1.00 (N=2)
4	0.22 (N=4)	1.00 (N=2)	-1.00 (N=2)	N/A
5	N/A	1.00 (N=2)	N/A	1.00 (N=2)

With respect to the low correlation in the benchmark group 4 at A1 level, it comes from one of the higher band scripts having the lower measure than the band one scripts. With that script excluded, the correlation would be 0.96. At B1 level for group 4 benchmarks, the script with the lower score band had the higher measure. The low correlation at B2 level for group 2 benchmarks also represents the reversal of the script score compared to the measure. In this case, the scripts were only 1 score point apart, hence it is unsurprising that such reversal may have occurred.

Script scores were also available for one SfL AO. As can be seen from the table below, some correlations between measures and script scores were noticeably lower, though this is to some extent unsurprising given that the scripts within each RQF level represented a narrow range of marks around the cut score, and hence a relatively narrow range of quality. This is a familiar pattern, also observed in Curcin and Black (2019) with GCSE scripts on and around grade boundaries.

Table 23 Script score-measure correlations – SfL scripts

RQF level	N scripts	Mark range	Score-measure correlation
E1	10	7	0.78
E2	3	5	0.57
E3	10	6	0.51
L1	10	7	0.12

Comparability between CEFR benchmarks

The chart below presents the distribution of benchmark script measures by benchmark group and the CEFR level of the scripts. There were 2 outlier benchmarks at A2 level, which were removed from analysis.

It can be seen that there is some variability in the perceived quality of scripts within each CEFR level. This was to be expected given that the scripts representing each level varied in terms of original marks/mark bands (for instance, with some scripts at the borderline and some secure within a level) or came from different examinations that might have somewhat different standards compared to others. Nevertheless, the scripts in higher score bands within their organisations tended to get higher measures than the borderline score band scripts. Furthermore, it can be seen that there are clear increases in quality between A1, A2 and B1 levels in particular, though there is more overlap between B1 and B2 levels.

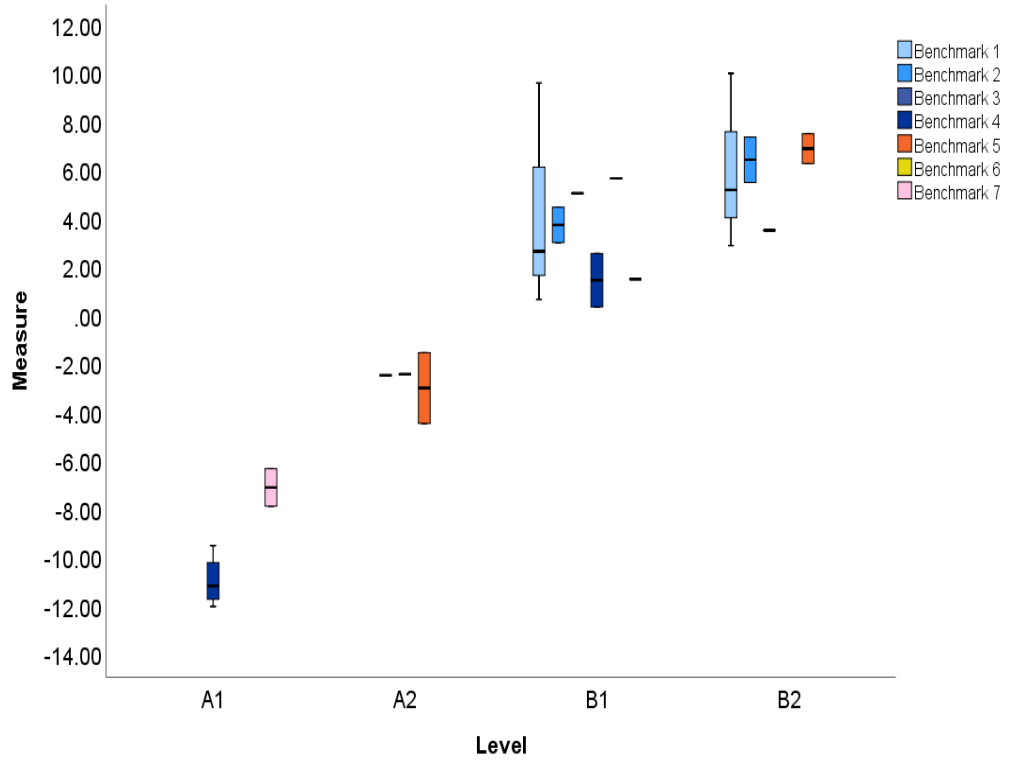


Figure 2 Comparability between CEFR benchmarks

This is further demonstrated in the table below, which shows median and mean logit measure of script quality by CEFR level. It can be seen that there is a clear distinction between the levels in terms of median and mean measure for all the scripts representing a level. In the remaining discussion, therefore, we grouped all the benchmark scripts representing a level into a single benchmark group, comparing these joint distributions by CEFR level with those of the SfL scripts.

Table 24 CEFR benchmark descriptive statistics by level

CEFR level	N	Median	Mean	Std. Deviation	Range	Minimum	Maximum
A1	6	-10.27	-9.74	2.22	5.69	-12.08	-6.39
A2	4	-2.51	-2.79	1.23	2.92	-4.52	-1.60
B1	10	2.76	3.49	2.77	9.25	0.29	9.54
B2	8	5.82	5.96	2.30	7.12	2.82	9.93

Borderline SfL script check

It was not always straightforward to ensure our SfL sample contained truly borderline pass grade scripts. This was particularly problematic where the pass grade was not defined in terms of numerical marks but rather in terms of binary criteria fulfilment.

We therefore requested a small sample of borderline fail scripts from all AOs, to check that the pass grade scripts in each AO's sample were reasonably close in quality to borderline fail scripts. Charts 3 to 6 display the measures that each AO's fail and pass scripts obtained, plotted next to each other for each RQF level.

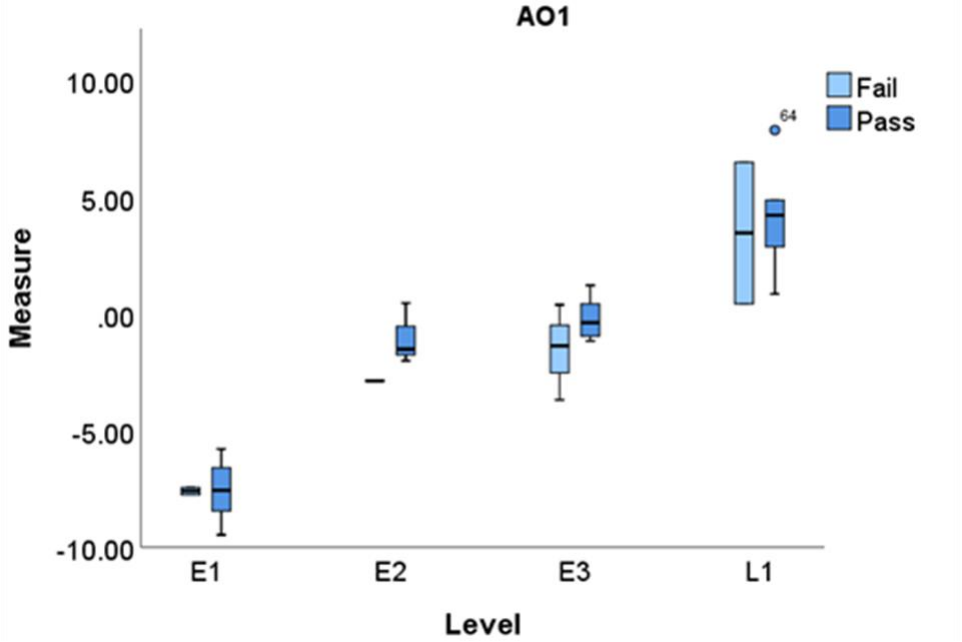


Figure 3 SfL sample borderline script check – AO1

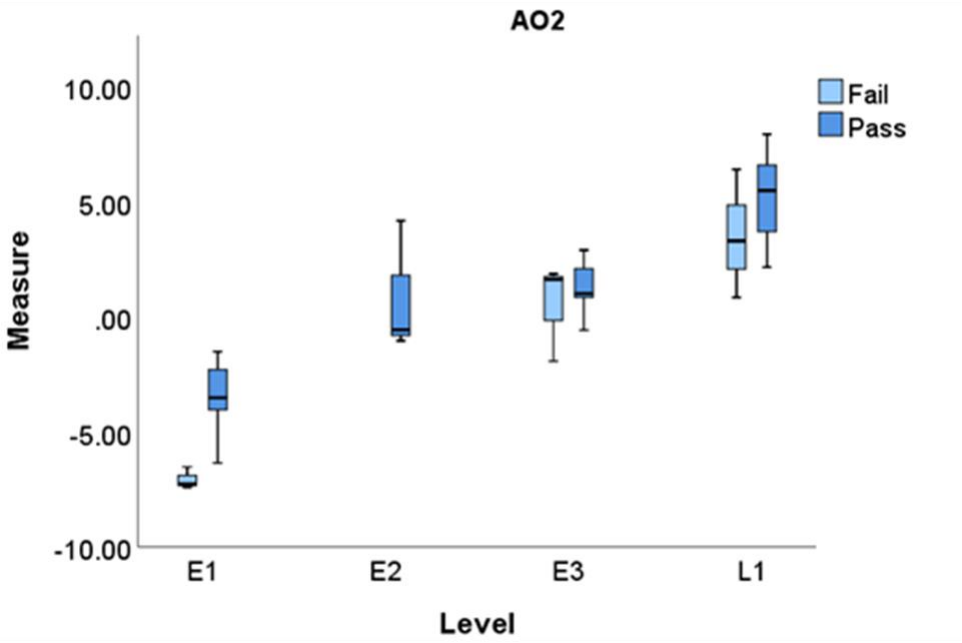


Figure 4 SfL sample borderline script check – AO2

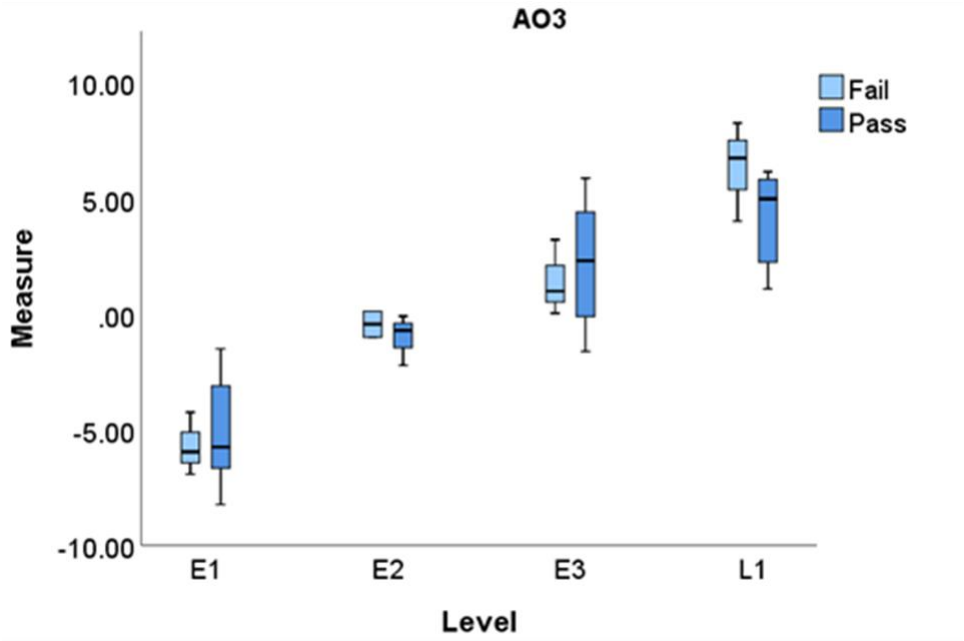


Figure 5 SfL sample borderline script check – AO3

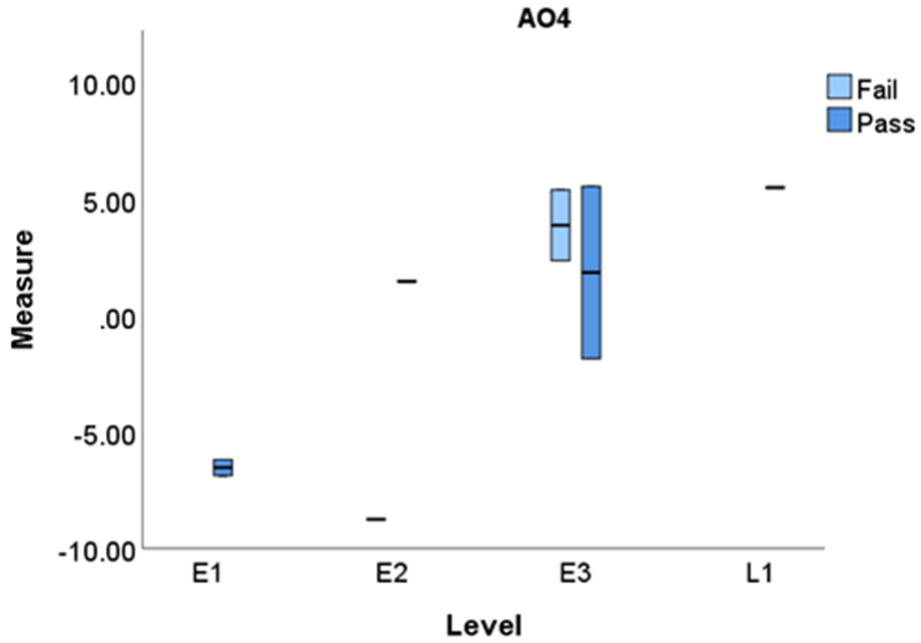


Figure 6 SfL sample borderline script check – AO4

It can be seen that in the majority of cases, where both fail and pass scripts were available, they were either fairly close in terms of their measures, or there was a degree of overlap, with some fail scripts even judged to be superior to some of the pass scripts (for instance, AO3 at L1). In the case of AO4, the data submitted by the AO was not always clear as to whether each script was borderline or not. Given the

difference between the fail and the pass script measure for this AO at Entry 2, it is likely that at least one of them was not borderline.

Overall, it can be concluded that at least some of the pass scripts that were submitted did represent borderline pass standard. For the analyses presented in the next 2 sections, we therefore only focus on the pass scripts.

Comparability between SfL AOs

One of the main questions in this phase of the study was to consider how comparable the performance standards of candidates at the pass grade threshold are between different AOs in our sample. Recall that AO4 provided a very small number of scripts, so their results are only included for tentative comparisons with the others.

Firstly, as can be seen from the Figure 7 and table 25, there is a fair bit of variability in perceived quality between the scripts from most AOs at each level. This partly just reflects the fact that not all the scripts that were submitted by each AO were borderline.

With respect to how the pass grade standards between different AOs relate to each other at each level, it can be observed that there is a fair bit of overlap between the script quality measures at the bottom end of each level, and particularly at Entry 2, Entry 3 and Level 1 (with the exception of AO4, for which we do not have sufficient data). Within that, AO2 does appear to involve a slightly higher performance standard at the pass grade at those levels. Overall, however, the amount of quality measure overlap at Entry 2, Entry 3 and Level 1 suggests that the standards at the pass grade borderline for the AOs in our sample are reasonably comparable.

There is somewhat more variability between the AOs at Entry 1, with AO1 and AO2 appearing to have markedly different performance standards at the pass grade threshold, with AO3 (and AO4) somewhere in between. Our curriculum review did flag some aspects of AO2 standards as potentially being too high in relation to the curriculum at Entry 1, followed by AO3, particularly with respect to high expected word counts for the level, and this could partly explain their apparently slightly higher standard compared to AO1.

Another notable finding from this exercise is that there is not a great deal of differentiation between Entry 2 and Entry 3 scripts in particular, irrespective of the AO, but also to some extent between Entry 3 and Level 1 across all of the AOs. This is discussed in more detail in the next section.

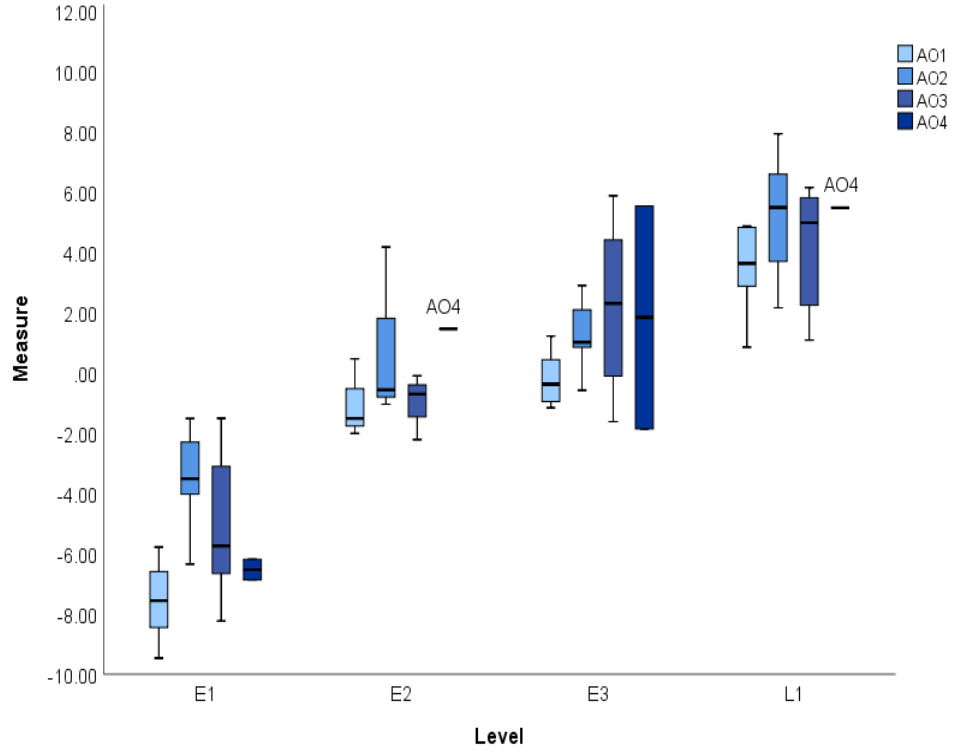


Figure 7 Comparability between SfL scripts by level

Table 25 Descriptive statistics for each AO script sample by level

Level	AO	N	Mean	Std. Deviation	Range	Min	Max
E1	AO1	7	-7.56	1.32	3.68	-9.47	-5.79
E1	AO2	7	-3.43	1.65	4.83	-6.35	-1.52
E1	AO3	6	-5.17	2.46	6.73	-8.24	-1.51
E1	AO4	2	-6.53	0.48	0.68	-6.87	-6.19
E2	AO1	3	-1.02	1.31	2.47	-2.01	0.46
E2	AO2	3	0.85	2.89	5.22	-1.05	4.17
E2	AO3	3	-1.01	1.09	2.12	-2.22	-0.1
E2	AO4	1	1.45	-	0	1.45	1.45
E3	AO1	7	-0.20	0.95	2.38	-1.16	1.22
E3	AO2	7	1.31	1.19	3.47	-0.58	2.89
E3	AO3	8	2.18	2.79	7.49	-1.62	5.87
E3	AO4	2	1.84	5.23	7.39	-1.86	5.53

L1	AO1	5	3.41	1.66	4.02	0.85	4.87
L1	AO2	7	5.16	2.18	5.78	2.15	7.93
L1	AO3	5	4.05	2.26	5.06	1.08	6.15
L1	AO4	1	5.47	-	0	5.47	5.47

Linking to the CEFR

In this section we present the results of the linking of the SfL performance standards to the CEFR. The CJ exercise also involved several ESOL Exemplars samples, and below we comment on the relationship between the SfL standards and those of ESOL Exemplars. The ESOL Exemplars show the work of learners who are mainly secure in a level rather than having a borderline performance.

The relationship between the quality measures (performance standards) of the SfL scripts (blue bars), the CEFR benchmarks (orange bar), and the ESOL Exemplars (yellow bar) is presented in the Figure 8[24]. The chart displays box and whisker plots of the script quality measures from the rank ordering exercise. On the chart, the borderline standard at each CEFR level is denoted by the dotted lines and the relevant CEFR level labels.

[24] The table with the descriptive statistics for each of the three script groups by level is given in appendix 6.

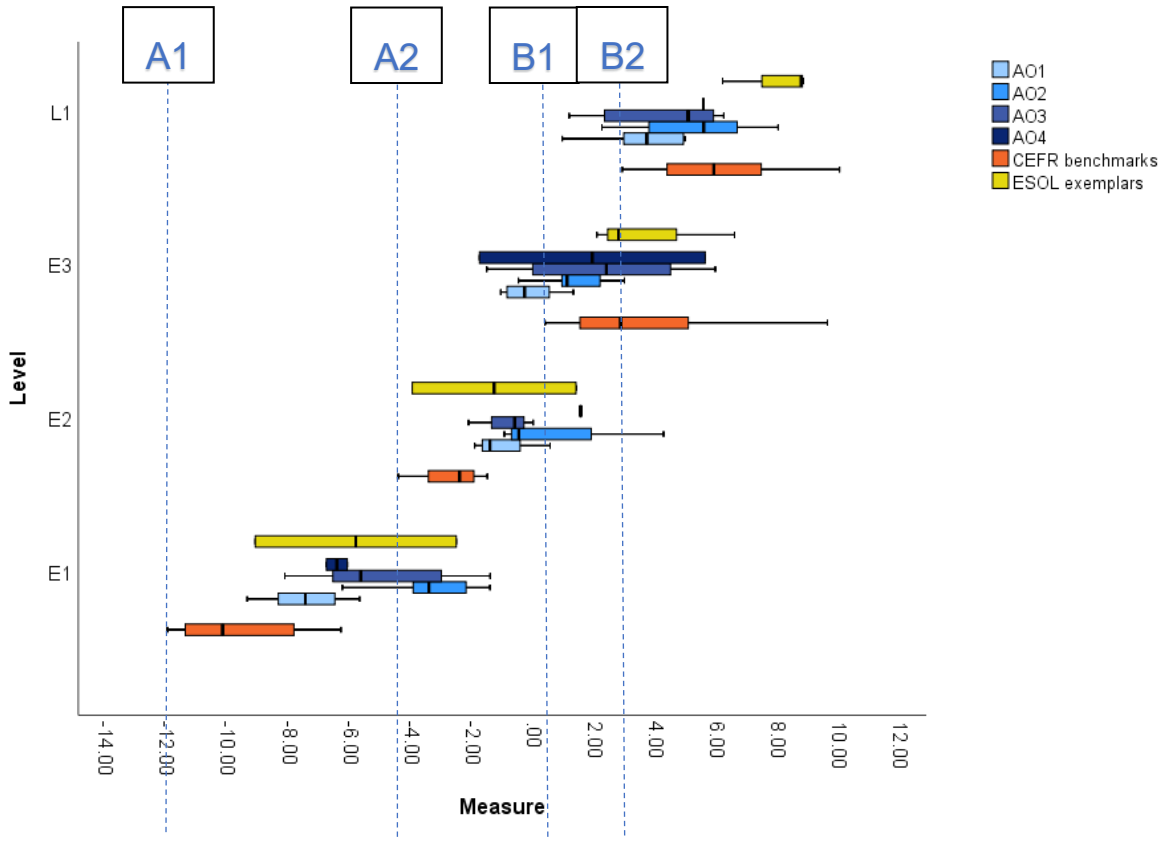


Figure 8 Box plots showing the SfL standards mapping onto the CEFR

At Entry 1, the borderline pass grade standard of SfL scripts is well within the A1 level range compared with the CEFR benchmarks, despite some differences between individual AOs (see previous section). In fact, several scripts appear to be of A2 level quality. Therefore, the pass grade standard at Entry 1 can be categorised as mid A1 to A1+. There is also good alignment with the standard of ESOL exemplars, which is also higher than threshold A1 CEFR level.

Similarly, at Entry 2, the pass grade standard seems to be well within the A2 CEFR range, at A2+ level, and somewhat higher than the standard of some ESOL Exemplars.

At Entry 3, it is notable that the SfL script quality is lower than threshold B1 CEFR level for a number of scripts in our sample from AO1 and AO3, as well for one script from AO2 and the only borderline script we had from AO4. In addition, nearly 70% of the SfL scripts in our sample across all the AOs were of lower quality than all except 2 CEFR benchmark scripts. There is also significant overlap between the quality measures of Entry 2 and Entry 3 SfL scripts from all AOs. This would suggest that at least some Entry 3 SfL borderline pass grade standards are more in line with A2+ across all AOs, with around 70% of the sample just in the low B1 level. At Entry 3, the ESOL Exemplars are of noticeably higher standard than the majority of SfL pass grade scripts. The ESOL exemplars straddle B1+/low B2 CEFR level.

At Level 1, the majority of the SfL scripts are at low B2 level, although several borderline scripts do not reach that level, and are more appropriately mapped to B1 mid or B1+ level. Again, the ESOL exemplars are higher in quality than the majority of the SfL scripts in our sample, and well within the B2 level.

Table 26 Relationship between SfL pass grade standard, ESOL exemplars and the CEFR

RQF level	SfL scripts (Borderline pass)	ESOL exemplars (Secure at level)
E1	Mid A1/A1+	A1+/low A2
E2	A2+	Mid A2
E3	A2+/low B1	B1+/low B2
L1	B1+/low B2	Mid B2/B2+

The findings from the CEFR mapping indicate that the SfL writing assessments from the 4 AOs in our sample are broadly comparable in terms of their CEFR level, despite some discrepancies at the pass grade borderline within A1 level in particular. The more significant issue than inter-AO comparability is the fact that there does not appear to be enough differentiation between different SfL levels at the pass grade threshold, particularly between Entry 2 and Entry 3, which can both be mapped onto A2+ or low B1 CEFR level for at least part of the sample across all AOs. Most of the SfL Entry 3 sample can be mapped to low B1 level at most.

This mapping also indicates that ESOL Exemplars, which exemplify secure performance within the RQF levels of the curriculum, map onto the mid to high A1, A2, B1 and B2 CEFR levels respectively. This suggests that the borderline performance manifestation should be somewhat lower within each CEFR level, but still map onto each level at about its threshold, rather than dropping below it, as seems to be the case with the SfL writing qualifications in our sample at Entry 3 and Level 1. The CEFR content mapping also indicated that the written assessments at Entry 1 and Entry 3 are at around A1/A1+ and B1/B1+ respectively, or rather, that they have the potential to provide performance evidence of those levels. It is another issue where the threshold standard is actually set on any specific assessment, as demonstrated by the Entry 3 pass grade threshold mapping onto A2+ rather than threshold B1 or higher for several borderline scripts.

The findings from this mapping largely confirm what was suggested in the curriculum review when the SfL qualification demands were considered against the curriculum and NSAL standards. Namely, that Entry 3 and Level 1 qualifications may not be sufficiently demanding given the demand implied in the curriculum and relative to the demand of the lower levels. The curriculum review also indicated that the demand of Entry 1 is mostly appropriate, if a little too high with some AOs. This is also

supported by the indicative relationship found here between Entry 1 (and Entry 2) SfL samples and ESOL Exemplars.

Therefore, the issue of lack of differentiation between Entry 2 and Entry 3 in particular appears to stem from the standards of Entry 3 likely being too low, threatening the progression of learners onto higher levels. This apparent lack of stretch at Entry 3 may be contributing to the perceived “large jump” in standards at Level 1, which was noted by our curriculum reviewers, as well as some stakeholders.

It is also worth pointing out that even though the Entry 1 qualifications may be compatible with mid-high A1 level according to the curriculum, the question remains whether this is suitable for an initial Entry level qualification. Given that Entry 1 pass grade standard appears to be noticeably higher than threshold A1 CEFR level, this suggests a potential need for separate qualifications at somewhat lower level (’pre--entry’), as the current GLH may not enable some learners to achieve the equivalent of mid-high A1 level when starting from the point of no English in SfL Entry 1 qualifications.

Some caveats

As with the CEFR mapping of assessment standards for reading comprehension, there are some caveats that need to be borne in mind when interpreting the CEFR mapping of the writing pass grade standards. These caveats had already been identified in the content mapping phase, and relate to where the assessments or the constructs assessed depart from the CEFR in relation to:

1. use of some assessment approaches and task types that may not be optimal for eliciting communicative language competence
2. approach to literacy assessment in relation to the requirement of a degree of accuracy in spelling and punctuation and an awareness of the writing process which are not expected at corresponding CEFR levels

The first category of discrepancies involves, for instance, limited use of sufficiently interactive writing tasks in some AOs’ assessments, occasional assessment of knowledge of grammar or spelling through non-functional gap-filling or selected response tasks within the writing component, which may be more appropriate for instructional use, and the approach of assessing within a single topic area across all components. These discrepancies represent a departure from what might be considered best practice in terms of action-oriented communicative approaches to some extent.

While these discrepancies do not entirely preclude the possibility of mapping to the CEFR, they do limit the interpretation of the indicative levels in terms of how much evidence they entail of learners’ communicative ability in relation to some of the above-mentioned aspects. It should be noted, however, that such issues were only

present in the minority of AOs reviewed. Most of those issues were noted by the subject experts in our study 3, as well as aspects that were at odds with the core curriculum.

The mismatches in category 2 make the CEFR content mapping fuzzier, reflecting the specific context of the SfL qualifications. For instance, some Entry 1 assessments, which are broadly at A1/A1+ CEFR level, appear to require a degree of orthographic accuracy (such as spelling or use of capitalisation) that do not appear until B1 level in the CEFR. In addition, in the CEFR, the process of writing is not addressed explicitly, whereas in Entry 3 and Level 1 SfL writing assessments, the following are very clearly specified and required within some assessment criteria (related to their link to the NSAL):

- the process of planning and drafting when writing
- awareness that some types of writing do not need formal planning while others such as important letters or college assignments do
- proofreading and correcting writing for grammar and spelling

While the CEFR does include reference to some of the above-mentioned aspects in relation to pragmatic and sociolinguistic appropriateness, as well as in its orthographic control scale, and under coherence and cohesion, these are not always explicitly related to writing only, and are couched in terms of achieving communicative aims rather than being the focus of assessment. Therefore, even though these aspects may have been credit-worthy in achieving the pass grades in SfL assessments, they may not have contributed in the same way to our judges' views of the relative quality of different scripts that they considered in terms of the CEFR description of what constitutes effective writing at different CEFR levels.

Finally, as noted earlier, the form-filling task, which was present in each AO's writing assessment at each level, may have contributed only to a limited extent to judges' overall view of each script's quality. This is partly because some of the student personal details that achieve some credit operationally had to be removed. However, given that those details achieve most credit at lower levels, which in fact map quite highly within A1 and A2 CEFR levels respectively, this suggests that not considering those parts of the forms was not detrimental for the overall perception of those scripts in relation to the CEFR benchmarks or ESOL Exemplars.

Limitations

There are several potential limitations to this research in terms of generalisability, primarily springing from its limited scope. All the individual studies are based on a subset of 4 SfL AOs and a subset of levels. Therefore, even though there is some room for generalisation (the 4 AOs in the sample account for a large proportion of SfL certifications annually, the qualifications at other levels tend to have similar design and measurement approaches within each AO), it is possible that there are aspects that were not observed in the current study.

In relation to the curriculum review study, it should be noted that the views reflected in our findings come from only 2 reviewers, and therefore may be to some extent skewed. We attempted to guard against this by recruiting reviewers with slightly different backgrounds in terms of their closeness to either the SfL tradition or CEFR/EFL tradition, so that they could provide some challenge to each other's views and interpretations having done the reviews independently of each other. Nevertheless, for any higher-stakes purpose such as informing actual reforms to these qualifications or the curriculum itself, it would be advisable to consider a broader set of expert views and include experts in language assessment design more directly.

Similarly, with reference to the procedure for carrying out the CEFR linking for reading comprehension tests, as with other situations 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 (Asch, 1951; Deutsch and Gerard, 1955), polarisation (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 in panel discussions than might have been the case in a different set-up. On the other hand, the purpose of panel 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 preferable to use 2 different panels of experts for each specification and explore the extent to which their independent results agreed and supported each other. In addition, there is a range of other potential standard-linking methods that some might consider more appropriate for higher stakes uses, and which could be used sif item functioning information about the relevant tests were available (for instance, the Angoff or the Bookmark method).

With respect to the CEFR linking study more generally, it is important to note that it was conducted as a research exercise, to consider issues related to comparability and compatibility of content, rather than as an official linking study. Given its purpose, it was considered sufficient to focus the linking on a subset of SfL

qualifications, components and levels at this point to serve as an example, but again, this limits the interpretation of the findings to those components and AOs. 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 components on which the linking was performed.

The CEFR linking of the writing components was additionally somewhat limited by availability of ideally suitable candidate scripts, as well as by the insufficient number of samples from one AO. The latter issue did not really allow clear conclusions about this AO's pass grade standards, and additionally reduces overall generalisability of our findings, even though the limited data from that AO largely accords with the patterns across the other AOs.

With respect to using the rank ordering method to create a common quality scale across the CEFR benchmarks and SfL scripts, while the exercise appears to have worked reasonably well for our research purposes, with largely consistent judgements and high scale separation reliability, it might be advisable to replicate or confirm these results using a more traditional standard-linking methodology, which would not need to rely on using CEFR benchmarks and would just use the CEFR scales. This would avoid challenges with ensuring that the CEFR benchmarks are sufficiently representative and of comparable borderline standards. On the other hand, such an exercise would require that the judges were far more thoroughly standardised as a group in using the CEFR scale categories, and could also suffer from individual judge severity/leniency effects and group dynamics effects if involving panel discussions.

Discussion

The purpose of this research was to explore and evaluate a sample of SfL qualifications in terms of their design and assessment properties, alignment with the core curriculum, inter-AO comparability and validity. This provides us with an understanding of the effectiveness of our current regulatory requirements for these qualifications in ensuring their optimal functioning and validity in relation to their purposes. In this section, we bring together the findings from our 4 studies and suggest some implications. The discussion is organised according to some key themes that emerged from one or more of our studies or from our background research.

A challenging assessment context

The SfL qualifications were developed as part of a suite of basic skills qualifications to foster and improve linguistic integration of different categories of (adult or over-16) migrants who are living in the UK. The SfL qualifications are meant to align with the core curriculum and the NSAL, and it should arguably be a matter of indifference to learners which AO's qualification they are taking in terms of how the results are interpreted and used. This requirement calls for a degree of comparability of standards between these qualifications in the challenging context of language assessment for migration and integration.

The challenge is primarily due to the sheer variety among the target cohort, in terms of literacy levels, language and educational backgrounds, life histories including trauma and other vulnerabilities, aspirations and needs within the UK society, which often calls for flexibility in delivery and assessment approaches, in a non-threatening environment. In the context of language assessment for migration/integration, it has been argued that where qualifications are mainly used or needed to enhance linguistic integration, and do not primarily serve high stakes immigration or other high stakes purposes, the flexibility and quality of course provision can be seen as equally if not more important than ensuring assessment standardisation (Rocca, Hamnes Carlsen & Deygers, ibid.; cf. O'Leary, 2008).

On the other hand, insofar as the results of these qualifications may be needed or used for progression to employment or further education, it would be important that a reasonable degree of assessment quality and validity, as well as a degree of comparability given the shared curriculum, is ensured so that the results can be recognised as trustworthy. Finding the right balance between addressing individual learner needs, contexts and different qualification purposes, while satisfying certain best-practice assessment principles, especially perhaps that of comparability, is likely to be a challenging task for both the AOs and the regulator. What is also

apparent, and discussed in the following sections, is that this challenging context of varying learner needs and stakeholder requirements may also create challenges around defining the purposes of SfL qualifications in the first place.

Issues of recognition

It is clear from the views expressed in our interviews with AO staff, as well as from the literature, that the SfL qualifications are seen by practitioners as having enormous value to learners and society as a whole. Our interview participants felt, however, that the potential of SfL qualifications to enable learner progression was not sufficiently recognised. In our interviews, as well as in the relevant literature, there was also a sense that these qualifications are undervalued in terms of funding. The AOs told us that this affected stakeholder attitudes towards the qualifications, including centres, employers, higher education institutions, and learners themselves.

AOs saw this lack of recognition of SfL qualifications as unjustified (this view is also apparent in the relevant literature), and all AOs emphasised the need to recognise the unique value of these qualifications for addressing ESOL learner needs in terms of developing their broad language skills, including literacy skills, from basic to advanced, while recognising their parity with similar qualifications such as FSE. The SfL qualifications were not seen as dissimilar in their enabling function to ESOL International qualifications, some of which are widely recognised, at competence levels similar to Entry 3, Level 1 or Level 2, by reputable higher education institutions as well as for visa purposes and job recruitment.

Multiple qualification purposes

Using Newton's (2017) perspectives framework for thinking about assessment purposes, based on our background research, it seems fairly clear that inasmuch as these qualifications are seen as valuable for linguistic integration of migrants and for removing barriers to accessing employment or education, some of their purposes can be seen from an **engagement perspective**, helping to ensure that learning actually occurs. Given their required alignment with the core curriculum, it appears that some of the purposes relate to the **expertise perspective**, providing evidence of a level of mastery or the acquisition of expertise of the specified content domain, as well as boosting confidence and motivation to build on skills and expand the expertise within the domain. It is also fairly clear from the literature that the SfL qualifications are seen as both socially and personally valuable in relation to these 2 perspectives. What is less clear, based on the issues with recognition of these qualifications for progression to employment or education, is how far they have (or should have) purposes related to **information perspective** that are valuable to

stakeholders beyond learners or qualification providers as the basis for making any decisions.

This could also be seen as a situation where a qualification justifiably has multiple (competing) purposes (cf. Newton, ibid.), but where there is not enough clarity about their relative importance to different stakeholders, and no simple answer to the question of why these qualifications are actually needed. This situation may indirectly impact on overall quality and appropriateness of provision and reduce the extent to which any of the intended purposes are truly fulfilled, potentially reducing the usefulness of these qualifications to users.

From the regulatory perspective, it would be important to seek clarity about the intended purposes of these qualifications and to clearly acknowledge and clarify whether there may be multiple purposes and their relative importance to stakeholders. This is necessary to enable coherent considerations regarding the most appropriate assessment models, as well as approaches to regulation, and any trade-offs that might need to be acknowledged or accepted, in order to ensure that the SfL qualifications are as fit for their purpose(s) as possible and regulated in the most appropriate way.

Assessment models

From our internal SfL assessment review we could see that a wide range of assessment models and approaches are used across the 4 AOs in our sample. Some of the differences in approaches appear to stem from a desire to offer flexibility of choice to learners and centres, so that different needs can be satisfied. The AOs in our interviews thought that the flexibility of the SfL qualification structure, its modularity and the variety of assessment models that can be used, were beneficial and justified by the needs and the wide range of contexts of learners and centres.

All our interviewees referred to the flexibilities offered in the context of speaking assessments in particular. Perhaps unsurprisingly, however, AO staff in our interviews suggested that the flexibilities offered to centres in terms of written task amendments are rarely taken up in the context of reading or writing assessments as they are resource intensive, although the AOs themselves offered some choice of context across different assessment versions in some models. In addition, the choices that centres make, for instance, in relation to internal vs. external assessment models, do not seem to always be tailored to the learners, and appear to be driven by centre needs (for example, availability of assessors), highlighting another area where there may be competing needs that may not always have solutions that satisfy them all.

Our assessment review also suggested that some of the individual AO approaches may not be entirely appropriate for the construct of assessment, while some may be remnants of historical approaches and regulatory frameworks such as the QCF, or may be poorly executed in practice. The interviews with the AOs suggested that some of the assessment design choices were indeed historical and not always reflective of best practice in assessment (for instance, choice and implementation of measurement models, weighting of certain curriculum aspects, atomistic assessment criteria), and AOs themselves were not always able to provide a clear rationale for them beyond citing original requirements that had not been updated. Therefore, it would be important to ensure that individual AO assessment models are of sufficient quality and sufficiently valid and justified for the purposes and construct of these qualifications.

Nature of curriculum alignment

The curriculum alignment review suggests that the SfL qualifications may not be sufficiently well aligned with the curriculum. This was primarily in terms of demand, aspects of coverage, and interpretation of the curriculum messages.

Based on the review, the demand of assessments and pass grade standards at Entry 1 were seen as broadly appropriate, if occasionally too high in relation to the curriculum standards. However, at Entry 3 and Level 1, the demand was deemed to be too low in most cases.

The overall quality and appropriateness of coverage and relative demand of the assessments seem to decline from Entry 1 to Level 1 for most of the AOs in our sample, with insufficient differentiation and stretch in the progress through the levels. This is a significant finding and highlights threats around appropriate candidate progression.

Furthermore, there was evidence of construct-irrelevant assessment approaches, and narrowing of the curriculum. For instance, it was noted that inordinate weight was sometimes given to content statements that involve enabling subskills (knowledge of alphabetical ordering, knowledge of names for text and grammar features), which are singled out in the curriculum for instruction rather than assessment purposes. In some cases, this was at the expense of assessing the global skills more thoroughly at the appropriate level (for instance, global and detailed understanding of text meaning). As another example, the focus on more academic and professional reading and writing, which is largely missing from the assessments, is present in the curriculum. It is not clear why it does not feature sufficiently in the assessments, given that some of the purposes of SfL qualifications, especially at Level 1 and above, are likely to involve progression to further study and employment.

In most cases it was not clear why such apparent misalignment may have occurred, as the curriculum statements and examples were deemed by the reviewers as sufficiently clear and indicative of the appropriate level, and the curriculum itself largely fit for purpose.

Comparability

A comparison of pass grade standards in Entry 3 reading comprehension assessments between the 4 AOs in our sample, carried out in relation to the CEFR, demonstrated that the pass grade standards are broadly comparable and at around threshold/low B1 level for 3 of the AOs. However, for one AO, the pass grade standard was mapped to A2+, and is thus lower than the rest. This is despite the overall profile and demand of the tests from the 4 AOs being reasonably aligned.

Similarly, some comparability issues were noted in our CEFR mapping study of pass grade standards in writing at Entry 1 level. However, the CEFR mapping also indicates that the SfL writing assessments from the 4 AOs are broadly comparable and all at the same CEFR level, despite some discrepancies at the pass grade borderline within A1 level. Therefore, these comparability issues are of somewhat less concern than those identified in reading comprehension.

In addition to these comparability issues that were directly identified between some AOs' grading standards, potential threats to comparability were identified that might stem from either inconsistency in assessment models (perhaps in particular the disparate measurement models and their implementations) or misalignment with the curriculum in terms of aspects and nature of coverage of the content domain. Given the issues around clarity of purposes of SfL qualifications, it would be important to clarify how much comparability of approaches and standards should be seen as a desirable given the relative nature and weight of qualification purposes.

CEFR alignment

The indicative CEFR content mapping produced in our studies suggests meaningful alignment of the SfL qualifications with the CEFR in many respects, and thus also with ESOL International qualifications.

The CEFR content mapping, which gives an indication of the extent of alignment between the content and construct of SfL qualifications and the CEFR and gives an indication of the most appropriate CEFR level that can be evidenced, placed Entry 1 components at A1/A1+ level, except for speaking, which was mapped to A1+/A2. The relatively high CEFR level for speaking may be unexpected for an Entry 1 qualification. Nevertheless, this was deemed appropriate for speaking in our curriculum mapping review, based on the standards implied in the curriculum.

Furthermore, our CEFR standard linking of writing components indicated that Entry 1 writing pass grade standard is also relatively high, at about A1+ level.

The CEFR content mapping placed Entry 3 components at B1/B1+ level. However, the results of our standard linking study for reading comprehension suggests that actual pass grade standards at Entry 3 are lower, in some cases at A2+, and in all cases closer to B1 threshold level than B1+. Similarly, the standard linking for the writing components indicated that the pass grade performance standard at Entry 3 is at A2+/low B1 at most, while Level 1 was mapped to B1+/low B2.

Our CEFR linking exercise for writing components also indicates that ESOL Exemplars, which exemplify secure performance within the RQF levels of the curriculum, map onto the mid to high A1, A2, B1 and B2 CEFR levels respectively. This suggests that the borderline performance manifestation should perhaps be somewhat lower within each CEFR level, but still map onto each level at about its threshold at least, rather than dropping below it, as seems to be the case with some samples of the SfL writing qualifications in our study at Entry 3 and Level 1 and with some reading comprehension assessments at Entry 3.

Perhaps the most significant finding from the CEFR standard linking study for writing is the fact that there does not appear to be enough differentiation between higher RQF levels of the SfL qualifications at the pass grade threshold, particularly between Entry 2 and Entry 3, which can both be mapped onto A2+ CEFR level. This is also the case to some extent between Entry 3 and Level 1 writing qualifications.

The findings from the CEFR standard linking largely confirm what was suggested in the curriculum review, namely, that Entry 3 and Level 1 SfL qualifications may not be sufficiently demanding given the demand implied in the curriculum and relative to the demand of the lower levels. This apparent lack of stretch at Entry 3 in particular may be contributing to the perceived “large jump” in standards at Level 1, which was noted by our curriculum reviewers, as well as some stakeholders.

It is worth pointing out that even though some Entry 1 components may be compatible with A1+ level according to the curriculum, the question remains whether this level is suitable for an initial Entry level language qualification. Given that Entry 1 pass grade standard appears to be significantly higher than threshold A1 CEFR level at least in writing, this may support stakeholder calls for separate qualifications at somewhat lower level (‘pre-entry’), as the current GLH may not enable some learners to achieve A1+ level when starting from the point of no English at Entry 1.

Assessment of literacy and other caveats to the interpretation of the CEFR linking

The CEFR content mapping highlighted the discrepancies with the CEFR in relation to assessment of literacy in SfL qualifications (such as, assessment of Latin alphabet knowledge, emphasis on orthographic accuracy at Entry 1, emphasis on awareness of the writing process at Entry 3). In the context of SfL qualifications, such requirements in the domain of literacy are related to their link to the NSAL, which are primarily targeted at learners for whom English is the first language.

The results of our Study 3, as well as the CEFR content mapping, suggest that perhaps such requirements may not be entirely appropriate at all language ability levels in the context of learning English as a second language, and particularly in the migration context. There are suggestions in the literature (ALTE, 2016) that in the migrant context, the focus on linguistic accuracy may not be as relevant as the ability to get the main idea across, at least at lower ability levels (up to B1). The CEFR, and the ESOL core curriculum to an extent, appear to recognise this more clearly than the NSAL. This potentially raises questions about whether mapping English as a second language qualifications, such as SfL, to standards intended for English as the first language qualifications such as NSAL is entirely appropriate.

The interpretation of the CEFR pass grade standard linking for reading comprehension tests is also limited to an extent by certain issues of item design, test design and potentially low reliability of some tests and use of some assessment approaches that do not optimally elicit evidence of reading comprehension competence. Similarly, there were issues in some writing assessments with approaches that are not sufficiently communicative or interactive. Such issues can also limit the interpretation of the CEFR linking, where assessment operationalisation rather than the scope of the specification or intended demand, constrains the nature of the skills elicited by the assessments and thus reduces overall validity of both the CEFR link and the assessment itself. Furthermore, the curriculum review has also indicated that the approaches that are not well aligned with the CEFR are also not sufficiently well aligned with the core curriculum, which is, like the CEFR, fundamentally communicative and functional in approach.

Potential improvements

In light of these findings, Ofqual (and other relevant stakeholders where appropriate, for instance, the DfE) should consider how the qualifications could be improved to address some of the issues identified. Some areas for improvement are listed below:

1. The range, relative weight, and interactions of SfL qualification purposes should be clarified. Given the complexity of SfL learner needs and the importance of engagement-related purposes of these qualifications, revisiting the trade-off between these and any information-related purposes would seem particularly important as a starting point. This should help to inform the choice of the most appropriate assessment models and any trade-offs between the extent of standardisation and flexibility in assessment design and delivery.
2. There are quality and validity issues in some individual approaches and assessment instruments, which would need to be addressed. In addition, appropriateness of approaches that may be historically motivated rather than meaningfully linked to qualification purposes and constructs should be reconsidered. Assessment approaches should be sufficiently justified in terms of validity and best assessment practice.
3. The way the curriculum is covered and interpreted in some assessments may need to be improved in order to reflect its intention and implied demand.
4. Insufficient comparability between Entry 3 reading comprehension pass grade standards should be addressed if comparability were deemed important for certain purposes of SfL qualifications. Wider checks on comparability (for instance, between other components, levels and AOs) may also be warranted depending on the extent of comparability required.
5. Differentiation and progression between SfL qualification levels should be improved to ensure optimal progression for learners.
6. Mapping SfL qualifications to the CEFR might clarify to stakeholders how the skills at each RQF level relate to the skills of ESOL International learners, who might be competing for the same jobs or education courses upon entry to this country and improve recognition of SfL qualifications for similar purposes.

References

AERA/APA/NCME (2014). Standards for Educational and Psychological Testing.

ALTE (2016). Language tests for access, integration and citizenship: An outline for policy makers. Cambridge: Association of Language Testers in Europe.

ALTE (2020). ALTE Principles of Good Practice.

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.

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.

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.

Bell, R., Plumb, N. & Marangozov, R. (2017). Integration not Demonisation. The final report of the All-Party Parliamentary Group on Social Integration's inquiry into the integration of immigrants.

Benton, T., Leech, T. & Hughes, S. (2020a). *Does comparative judgement of scripts provide an effective means of maintaining standards in mathematics?* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Benton, T., Cunningham, E., Hughes, S. & Leech, T. (2020b). *Comparing the simplified pairs method of standard maintaining to statistical equating.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

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.

Black, B., He, Q. & Holmes, S. (2017). *Vocational and Technical Qualifications: Assessment Functioning of external assessments. An overview of the functioning of assessments in 27 qualifications and 49 units,* (Ofqual/17/6319). Coventry: Ofqual.

Black, B., He, Q. & Morin, C. (2019). *Vocational and Technical Qualifications: assessment functioning of examinations in performance table qualifications. An overview of 20 Level 3 assessments*, (Ofqual/19/6501). Coventry: Ofqual.

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

Bradley R.A. & Terry M.E. (1952). Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons. *Biometrika*, 39, 324–45.

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, UK: Qualifications and Curriculum Authority.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3:2, 77–101.

Brunfaut, T., & Harding, L. (2013). *Linking the GEPT Listening Test to the Common European Framework of Reference.* (LTTC-GEPT Research Report RG-05.)

Cambridge Assessment English (2016a). Cambridge English Key. Handbook for teachers for exams from 2016. Cambridge: Cambridge Assessment English.

Cambridge Assessment English (2016b). Cambridge English Preliminary. Handbook for teachers for exams from 2016. Cambridge: Cambridge Assessment English.

Cambridge Assessment English (2019a). Cambridge English Qualifications A2 Key. Handbook for teachers for exams from 2020. Cambridge: Cambridge Assessment English.

Cambridge Assessment English (2019b). Cambridge English Qualifications B1 Preliminary. Handbook for teachers for exams from 2020. Cambridge: Cambridge Assessment English.

Cambridge Assessment English (2019c). Cambridge English Qualifications B2 First. Handbook for teachers. Cambridge: Cambridge Assessment English.

Carter, B., & Bathmaker, A. (2017). Prioritising progression over proficiency: limitations of teacher-based assessment within technician-level vocational education, *Journal of Further and Higher Education*, 41:4, 460–474.

Casey, L. (2016). The Casey review. A review into opportunity and integration..

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). *Rasch Models in Health*. London, UK: Wiley & Sons.

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.

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

Council of Europe (2009). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual.

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Strasbourg: Council of Europe Publishing.

Curcin, M. & Black, B. (2019). *Investigating standards in GCSE French, German and Spanish through the lens of the CEFR,* (Ofqual/19/6559/1). Coventry, UK: Ofqual.

Curcin, M., Howard, E., Sully, K. & Black, B. (2019). *Improving awarding: 2018/2019 pilots*, (Ofqual/19/6575). Coventry, UK: Ofqual.

De Jong, J. H. A. L. (2009). *Unwarranted claims about CEF alignment of some international English language tests.* Paper presented at EALTA, Turku, Finland.

Department for Education and Skills (2000). *Breaking the Language Barriers.* London: DfES.

Department for Education and Skills (2001). *Adult ESOL Core Curriculum.* London: DfES.

Department for Education and Skills (2003). Pathways to Proficiency. The Alignment of Language Proficiency Scales for assessing competence in English language. London: DfES.

Department for Education and Skills (2004). *ESOL Exemplars*. London: DfES.

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.

Ecclestone, K. (2002). *Leaner Autonomy in Post-16 Education: The Politics and Practice of Formative Assessment*. London: Routledge Falmer.

Engelhard, G., Jr., Kobrin, J. L., & 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.

Foster, D. & Bolton, P. (2018). *Adult ESOL in England*. Commons Library briefing paper 7905.

Garland, P. (1998). Assessment in GNVQs: learning the hard way. *Research in Post-Compulsory Education*, 3:3, 329–344.

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.

Higton, J., Sandhu, J., Stutz, A., Patel, R., Choudhoury, A. & Richards, S. (2019). English for speakers of other languages: Access and progression.

HM Government (2019). *Integrated Communities Strategy Green Paper. Summary of consultation responses and Government response.*

Johnson, M. (2008). Exploring assessor consistency in a Health and Social Care qualification using a sociocultural perspective, *Journal of Vocational Education and Training*, 60:2, 173–187.

Jones, N. (2009). A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting. *Research Notes,* 37, 6–9.

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.

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

Kings, P. & Casey, H. (2014). ESOL Qualifications and funding in 2014: Issues for consideration. NRDC report for The Association of Colleges.

Learning and Work Institute (2020). *ESOL Learners' progression to Functional Skills*.

Little, D. (2012). *The linguistic integration of adult migrants and the Common European Framework of Reference for Languages (CEFR)*. Council of Europe: Language Policy Unit.

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

Lockyer, C. & Cadwallader, S. (2020). Internal assessment in existing national technical and vocational qualifications, (Ofqual/20/6594). Coventry: Ofqual.

Moscovici, S., & Zavalloni, M. (1969). The group as the polarizer of attitudes. Journal of Personality and Social Psychology, 12, 125–135.

Newton, P. (2017). There Is More to Educational Measurement Than Measuring: The Importance of Embracing Purpose Pluralism. *Educational Measurement: Issues and Practice,* 36:2, 5–15.

Newton, P. (2018). *Grading Vocational & Technical Qualifications,* (Ofqual/18/6441/3). Coventry: Ofqual.

Newton, P. & Lockyer, C. (2022). How 'CASLO' Qualifications Work, (Ofqual/22/6895). Coventry: Ofqual.

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

Ofqual (2008). *Regulatory arrangements for the Qualifications and Credit Framework*. Coventry: Ofqual.

Ofqual (2011). *Criteria for English for Speakers of other Languages (ESOL) Qualifications.* Coventry: Ofqual.

Ofqual (2019). *Corporate plan 2019-2022.* Coventry: Ofqual.

O'Leary, D. (2008). A Common Language: making English work for London. London: Demos.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24:2, 193–202.

QCA (2000). National standards for adult literacy.

QCA (2009). Evaluation of ESOL Skills for Life Qualifications.

Robitzsch, A. (2019). *Sirt: Supplementary item response theory models*. R package version 3.7–40.

Rocca, L., Hamnes Carlsen, C. & Deygers, B. (2020). Linguistic integration of adult migrants: Requirements and learning opportunities. Report on the 2018 Council of Europe and ALTE survey on language and knowledge of society policies for migrants. Strasbourg: Council of Europe Publishing.

Roden, J. & Osmaston, M. (2021). *ESOL Learners and Functional Skills English.* Report commissioned by Education & Training Foundation.

Rolfe, H. & Stevenson, A. (2021). *Migration and English Language Learning after Brexit*. Leicester: National Learning and Work Institute.

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.

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

Torrance, H., H. Colley, D. Garratt, J. Jarvis, H. Piper, K. Ecclestone, and D. James. (2005). *The Impact of Different Modes of Assessment on Achievement and Progress in the Learning and Skills Sector (London, LSDA for the LSRC)*. London: The Learning and Skills Research Centre.

Turner, H. L., van Etten, J., Firth, D. & Kosmidis, I. (2020). Modelling Rankings in R: The PlackettLuce Package. *Computational Statistics*, 35, 1027–1057.

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*, 26:5, 541–562.

Wahlberg, M., & Gleeson, D. (2003). 'Doing the business': paradox and irony in vocational education – GNVQ business studies as a case in point. *Journal of Vocational Education and Training*, 55:4, 423–446.

West Midlands Combined Authority (WMCA) (2019). Unlocking Potential - Making Sense of ESOL in the Region.

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

Appendix 1: The CEFR global scale

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.

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.

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.

Appendix 2: Curriculum alignment review questions

AO-specific questions

1. Does the demand of the assessments appear broadly appropriate according to the standards in the Adult ESOL Core Curriculum and the National Standards for Adult Literacy at this level?
2. Is the coverage of the content in the assessments, alongside the minimum pass grade criteria, sufficient to evidence mastery of the curriculum at this level?
3. Do the assessments assess an appropriate sample of the vocabulary (including spelling), types of words, grammatical forms and communicative functions indicated in the curriculum for this level?
4. Are there any prominent standards and/or curriculum statements that are under-assessed (for example, that the assessments do not elicit enough evidence of from candidates)?
5. Is there too much emphasis on some aspects of the curriculum in the assessments?
6. Are the approaches used to assess the curriculum statements generally valid and appropriate for the nature of the content they are intended to assess?
7. Do the approaches used to assess the curriculum statements reflect the intention of the curriculum?
8. Are topics and contexts in the assessments appropriate and as intended by the curriculum at this level?
9. Are stimulus materials (for instance, reading texts, writing or speaking prompts) appropriate and as intended by the curriculum at this level?
10. Is the reading load and level of the language in the rubrics and questions appropriate at this level?
11. Do the questions/tasks test subject content in a way that is construct relevant?
12. In relation to the writing and speaking/listening components, are the assessment criteria/level descriptors clear and designed so they can be applied easily and consistently?

13. Are there any questions, tasks or stimulus materials that exhibit bias or may be inaccessible to some groups of learners such as learners with protected characteristics, or learners of particular ethnicity?

General questions

1. Is the way the curriculum statements are assessed across the 4 AOs broadly similar and, where different approaches are used, are these equally valid? If not, what are the most notable discrepancies between AOs and which curriculum statements do these relate to?
2. Is the standard implied by the coverage and the demand of assessments, as well as the pass grade criteria, broadly comparable across the 4 AOs?
3. Are there aspects of the curriculum at this level that might benefit from updating to make it more relevant to life in the UK in 2021 and beyond, or that you think should be changed or removed?
4. Is there anything that is missing from the current curriculum at this level that you think it would be helpful to add?
5. Are there aspects of the curriculum at this level that would benefit from more explicit alignment to the CEFR?

Appendix 3: Detailed CEFR content mapping tables

Entry 1

AO1 final estimation summary

Reception overall

Overall oral comprehension	Overall reading comprehension
A2	A1+

Communicative language competence in reception

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Reading comprehension	A1+/A2	A1+	A2	A1+/A2
Oral comprehension	A1+/A2	A1+/A2	A2	A1+/A2

Production overall

Overall oral production	Overall written production
A2	A1+

Communicative language competence in production

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral production	A2	A1+	A1+/A2	A2
Written production	A2	A1+	A1+/A2	A2

Interaction overall

Overall oral interaction	Overall written interaction
A2	A1+

Communicative language competence in interaction

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral interaction	A2	A1+	A1+/A2	A2
Written interaction	A2	A1+	A1+/A2	A2

AO2 final estimation summary

Reception overall

Overall oral comprehension	Overall reading comprehension
N/A	A1+

Communicative language competence in reception

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Reading comprehension	A1	N/A	A1+	A1+
Oral comprehension	N/A	N/A	N/A	N/A

Production overall

Overall oral production	Overall written production
N/A	A1

Communicative language competence in production

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral production	N/A	N/A	N/A	N/A
Written production	A1	A1	N/A	N/A

Interaction overall

Overall oral interaction	Overall written interaction
A2	A1

Communicative language competence in interaction

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral interaction	A2	A2	A2	A2
Written interaction	A1	A1	A1	N/A

AO3 final estimation summary

Reception overall

Overall oral comprehension	Overall reading comprehension
A1+/A2 low	A1+

Communicative language competence in reception

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Reading comprehension	A1+	A1+	A1+	A1+
Oral comprehension	A1+	A1+	A1+	A1+

Production overall

Overall oral production	Overall written production
N/A	A1/A2

Communicative language competence in production

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral production	N/A	N/A	N/A	N/A
Written production	A1/A2	A1/A2	A1/A2	A1/A2

Interaction overall

Overall oral interaction	Overall written interaction
A2+	A2

Communicative language competence in interaction

	Linguistic	**Socio-linguistic**	**Pragmatic**	**Strategic**
Oral interaction	A2/A2+	A2/A2+	A2/A2+	A2/A2+
Written interaction	A2	A2	A2	A2

AO4 final estimation summary

Reception overall

Overall oral comprehension	Overall reading comprehension
A1+	A1+/A2

Communicative language competence in reception

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Reading comprehension	A1+/A2	N/A	A2	A1+
Oral comprehension	A1+	n/a	A2	A1+

Production overall

Overall oral production	Overall written production
A2	A1

Communicative language competence in production

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral production	A1+/A2	A1+	A1+	A1
Written production	A1+/A2	A1	A1	A1

Interaction overall

Overall oral interaction	Overall written interaction
A1+/A2	A1

Communicative language competence in interaction

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral interaction	A1+	A1	A1+	A1
Written interaction	A1	A1	A1	A1

Entry 3

AO1 final estimation summary

Reception overall

Overall oral comprehension	Overall reading comprehension
B1	B1

Communicative language competence in reception

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Reading comprehension	B1	A2	B1	B1
Oral comprehension	B1	A2	B1	B1

Production overall

Overall oral production	Overall written production
B1	B1/B1+

Communicative language competence in production

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral production	B1	B1	B1	B1
Written production	B1	B1	B1	B1

Interaction overall

Overall oral interaction	Overall written interaction
B1/B1+	B1/B1+

Communicative language competence in interaction

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral interaction	B1	B1	B1	B1
Written interaction	B1	B1	B1	B1

Mediation

B1/B1+

AO2 final estimation summary

Reception overall

Overall oral comprehension	Overall reading comprehension
N/A	B1+

Communicative language competence in reception

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Reading comprehension	B1	B1	B1	B1
Oral comprehension	N/A	N/A	N/A	N/A

Production overall

Overall oral production	Overall written production
B1	B1

Communicative language competence in production

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral production	B1	B1	N/A	N/A
Written production	B1+	B2	B1	B1

Interaction overall

Overall oral interaction	Overall written interaction
B1+/B2	B1

Communicative language competence in interaction

	Linguistic	**Socio-linguistic**	**Pragmatic**	**Strategic**
Oral interaction	B2	B1	B1	B2
Written interaction	B1+	B2	B1	B1

AO3 final estimation summary

Reception overall

Overall oral comprehension	Overall reading comprehension
B1	B1+

Communicative language competence in reception

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Reading comprehension	B1	N/A	B1	B1+
Oral comprehension	B1+	N/A	B1	B1

Production overall

Overall oral production	Overall written production
B1	B1

Communicative language competence in production

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral production	B1	B1	B1	B1
Written production	B1+	B1	B1	B1

Interaction overall

Overall oral interaction	Overall written interaction
B1+	A2/B1

Communicative language competence in interaction

	Linguistic	**Socio-linguistic**	**Pragmatic**	**Strategic**
Oral interaction	B1+	B1+	B1+	B1+
Written interaction	B1+	B1	B1	B1

AO4 final estimation summary

Reception overall

Overall oral comprehension	Overall reading comprehension
B1	B1/B1+

Communicative language competence in reception

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Reading comprehension	B1	B1	B1	B1
Oral comprehension	B1	B1	B1	B1

Production overall

Overall oral production	Overall written production
B1	B1

Communicative language competence in production

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral production	B1	B1	B1	B1
Written production	B1	B1	B1	B1

Interaction overall

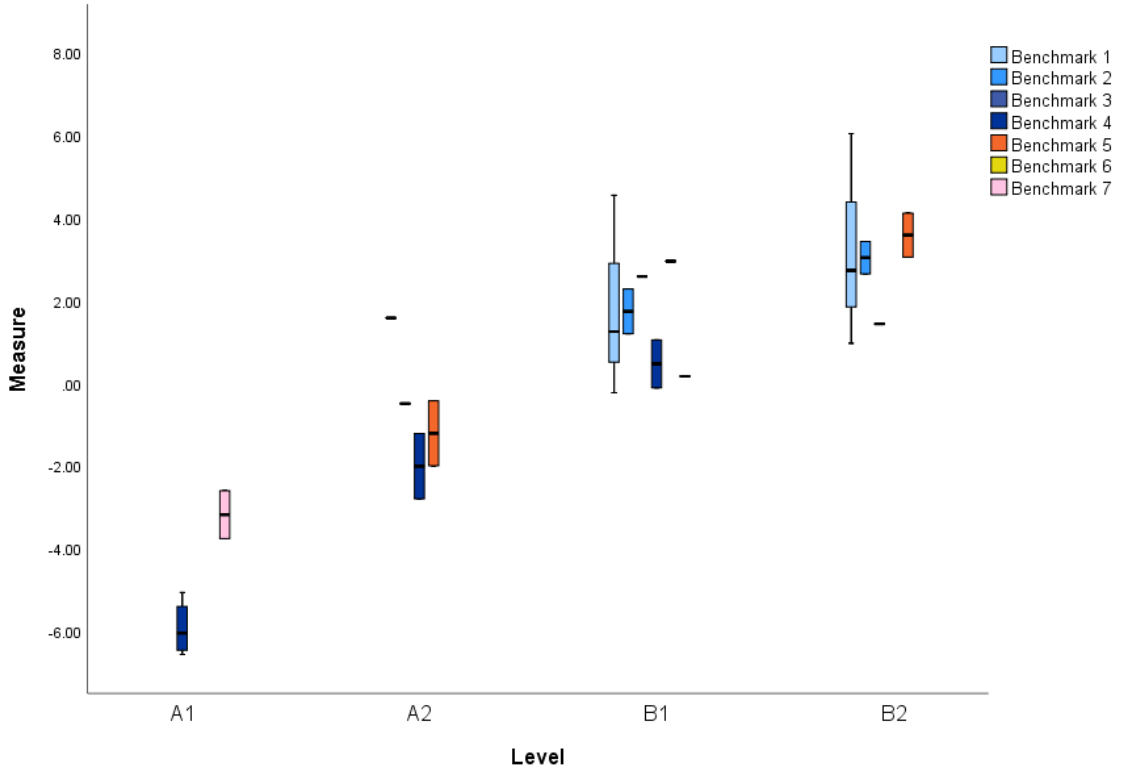
Overall oral interaction	Overall written interaction
B1	A2+/B1

Communicative language competence in interaction

	Linguistic	Socio-linguistic	Pragmatic	Strategic
Oral interaction	B1	B1	B1	B1
Written interaction	B1	B1	B1	B1

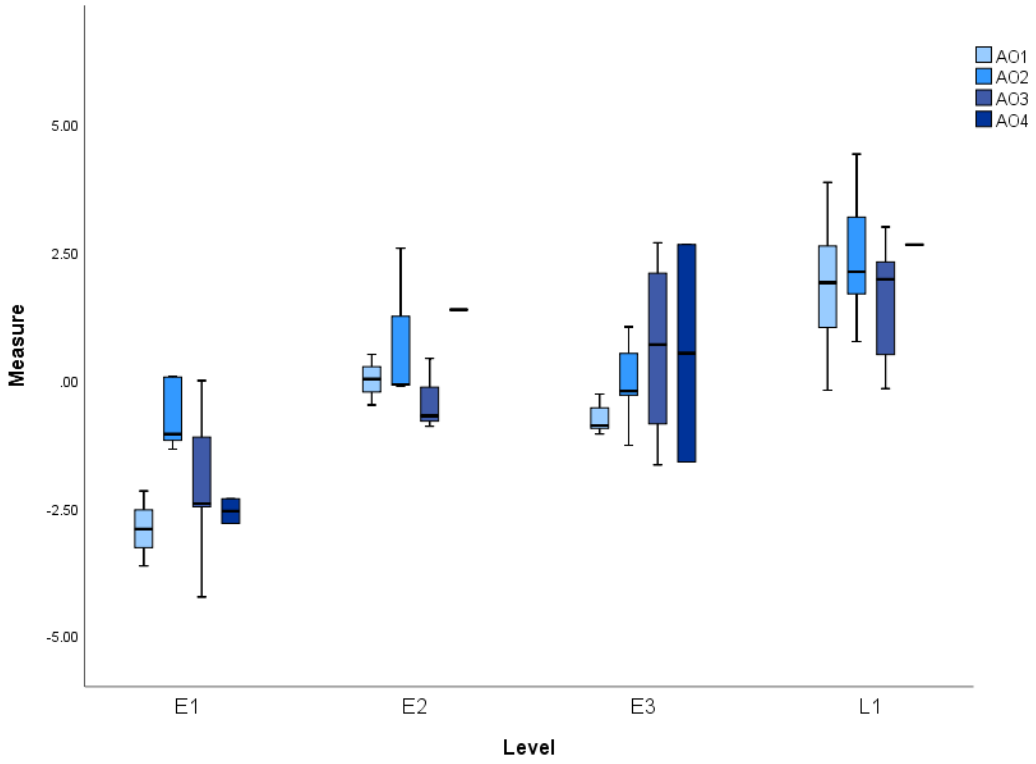
Appendix 4: PL-based analysis

Comparability between CEFR benchmarks



CEFR level	N	Median	Mean	Std. Deviation	Range	Min	Max
A1	8	-4.85	-4.42	2.01	5.89	-6.56	-0.67
A2	8	-0.85	-0.72	1.48	4.37	-2.8	1.58
B1	13	1.25	1.64	1.44	4.77	-0.23	4.54
B2	11	3.05	3.23	1.40	5.07	0.96	6.04

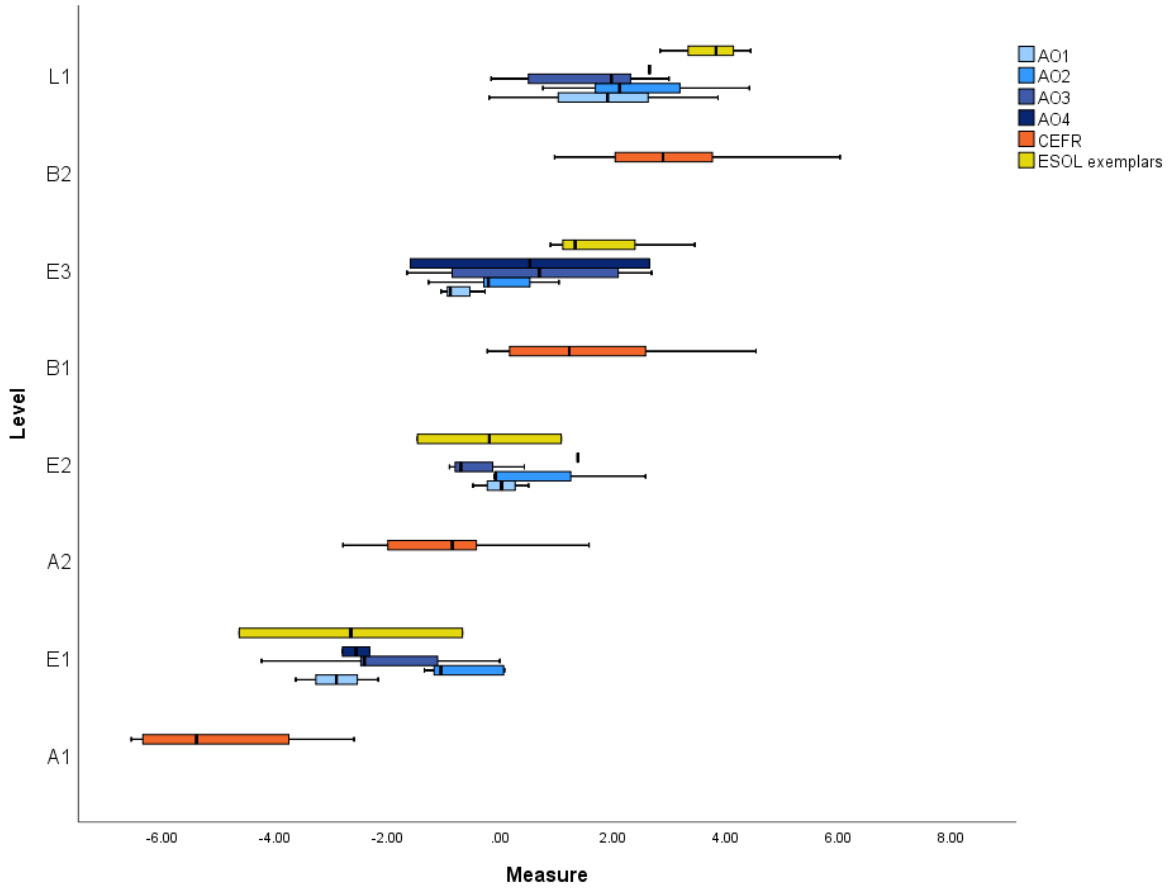
Comparability between SfL AOs



Level	AO	N	Mean	Std. Deviation	Range	Min	Max
E1	AO1	6	-2.91	0.53	1.47	-3.63	-2.17
	AO2	6	-0.75	0.64	1.42	-1.35	0.07

	AO3	6	-2.11	1.43	4.23	-4.24	-0.01
	AO4	2	-2.56	0.34	0.48	-2.81	-2.32
E2	AO1	3	0.01	0.49	0.99	-0.49	0.5
	AO2	3	0.79	1.55	2.7	-0.11	2.58
	AO3	3	-0.39	0.72	1.33	-0.9	0.43
	AO4	1	1.38	.	0	1.38	1.38
E3	AO1	6	-0.76	0.30	0.78	-1.05	-0.27
	AO2	7	0.00	0.80	2.32	-1.28	1.04
	AO3	8	0.61	1.70	4.35	-1.66	2.69
	AO4	2	0.53	3.01	4.25	-1.6	2.65
E4	AO1	6	1.86	1.41	4.07	-0.2	3.87
	AO2	7	2.44	1.33	3.67	0.75	4.42
	AO3	5	1.52	1.31	3.16	-0.17	3
	AO4	1	2.65	.	0	2.65	2.65

Linking to the CEFR



Appendix 5: Judge and script statistics

Judge statistics

Judge ID	Infit	Outfit
J01	0.62	0.25
J02	0.69	0.32
J03	0.89	0.48
J04	0.68	0.34
J05	0.83	0.35
J06	1.19	0.77
J07	1.03	0.58
J08	0.99	0.59
J09	0.76	0.35
J10	0.84	0.42
J11	0.57	0.24

Script statistics

Script	Measure	Measure SE	Infit	Outfit	Level	Chosen	Not chosen
155.pdf	9.93	1.90	0.09	0.01	B2	24	0
153.pdf	9.54	1.88	0.06	0.01	B1	23	0
149.pdf	8.76	0.99	0.67	0.16	L1	22	1
151.pdf	8.68	1.00	0.83	0.22	L1	21	1
215.pdf	8.25	0.83	0.89	0.36	L1	22	2
244.pdf	7.93	1.04	0.19	0.06	L1	22	2
272.pdf	7.87	0.73	1.09	0.47	L1	20	3
245.pdf	7.62	0.78	0.49	0.16	L1	20	4
123.pdf	7.44	0.66	0.95	0.54	B2	20	4
304.pdf	7.30	0.79	1.19	0.73	B2	19	4

214.pdf	6.74	0.74	0.48	0.18	L1	19	5
148.pdf	6.50	0.83	0.76	0.24	E3	21	3
274.pdf	6.49	0.81	1.70	0.64	L1	20	4
251.pdf	6.40	0.61	0.73	0.37	L1	18	6
110.pdf	6.21	0.67	0.75	0.33	B2	19	5
209.pdf	6.15	0.65	0.80	0.41	L1	14	7
150.pdf	6.11	0.72	0.68	0.26	L1	18	6
195.pdf	5.87	0.63	0.77	0.36	E3	16	8
212.pdf	5.81	0.61	1.66	0.99	L1	18	6
120.pdf	5.59	0.69	0.92	0.83	B1	19	5
247.pdf	5.56	0.67	0.75	0.33	L1	15	9
408.pdf	5.53	0.63	0.53	0.25	E3	16	8
246.pdf	5.49	0.68	1.09	1.16	L1	17	6
409.pdf	5.47	0.76	0.62	0.22	L1	17	7
305.pdf	5.43	0.63	1.01	0.51	B2	16	8
405.pdf	5.38	0.72	0.53	0.25	E3	16	8
157.pdf	5.12	0.71	0.97	0.38	B2	19	5
119.pdf	4.98	0.64	0.71	0.36	B1	18	6
213.pdf	4.98	0.72	0.75	0.48	L1	16	8
268.pdf	4.87	0.73	0.61	0.23	L1	19	5
270.pdf	4.82	0.72	1.36	1.05	L1	17	7
199.pdf	4.61	0.60	0.85	0.49	E3	17	7
242.pdf	4.51	0.57	0.59	0.35	L1	13	11
302.pdf	4.40	0.62	1.20	0.88	B1	17	7
196.pdf	4.20	0.66	0.67	0.33	E3	18	6
231.pdf	4.17	1.05	1.43	0.28	E2	22	2
217.pdf	4.01	0.65	1.10	0.54	L1	13	11
200.pdf	3.80	0.61	0.52	0.30	E3	11	13
269.pdf	3.63	0.68	0.36	0.18	L1	11	13
122.pdf	3.45	0.63	0.56	0.28	B2	14	10

250.pdf	3.30	0.67	0.48	0.23	L1	15	9
203.pdf	3.21	0.70	0.50	0.22	E3	15	8
303.pdf	2.94	0.60	0.93	0.61	B1	15	9
300.pdf	2.93	0.81	1.14	0.38	A2	20	4
237.pdf	2.89	0.74	1.24	0.96	E3	11	13
243.pdf	2.89	0.72	1.67	1.43	L1	10	14
271.pdf	2.87	0.63	1.05	0.70	L1	10	14
156.pdf	2.82	0.70	0.71	0.30	B2	10	14
238.pdf	2.69	0.69	0.35	0.17	E3	13	11
146.pdf	2.69	0.62	1.06	0.97	E3	13	10
152.pdf	2.58	0.59	0.69	0.39	B1	16	8
136.pdf	2.49	0.77	1.61	1.36	B1	9	14
407.pdf	2.35	0.75	1.24	1.22	E3	12	12
207.pdf	2.24	0.74	0.24	0.12	L1	9	15
248.pdf	2.15	0.70	0.65	0.39	L1	8	16
147.pdf	1.99	0.64	1.09	0.59	E3	15	9
239.pdf	1.87	0.56	0.76	0.49	E3	12	12
240.pdf	1.62	0.69	0.59	0.26	E3	8	15
236.pdf	1.49	0.56	1.07	0.70	E3	8	16
403.pdf	1.45	1.07	0.28	0.06	E2	22	2
125.pdf	1.43	0.73	0.47	0.18	B1	7	17
143.pdf	1.31	0.81	0.55	0.16	E2	18	6
171.pdf	1.22	0.55	0.67	0.42	E3	9	14
210.pdf	1.08	0.76	0.52	0.18	L1	6	18
234.pdf	1.01	0.64	0.93	0.60	E3	8	16
202.pdf	0.98	0.70	0.51	0.21	E3	7	17
266.pdf	0.95	0.69	0.70	0.29	E3	4	20
233.pdf	0.86	0.59	0.88	0.66	E3	7	16
176.pdf	0.85	0.68	0.83	0.37	L1	7	17
249.pdf	0.84	0.75	1.30	0.83	L1	5	18

235.pdf	0.83	0.72	1.21	0.77	E3	5	19
193.pdf	0.80	0.62	0.81	0.40	E3	9	15
201.pdf	0.73	0.64	0.99	0.50	E3	8	16
154.pdf	0.60	0.80	1.71	0.89	B1	7	17
164.pdf	0.46	0.67	0.97	0.54	E2	17	7
273.pdf	0.43	0.73	0.85	0.65	L1	3	21
175.pdf	0.38	0.77	1.14	0.52	E3	5	18
139.pdf	0.29	0.60	0.49	0.27	B1	6	18
189.pdf	0.12	0.80	0.51	0.16	E2	19	5
204.pdf	0.04	0.75	0.99	0.43	E3	6	18
167.pdf	-0.08	0.72	0.71	0.28	E3	6	18
191.pdf	-0.10	0.68	0.47	0.20	E2	15	9
168.pdf	-0.38	0.62	0.76	0.41	E3	7	17
230.pdf	-0.57	0.68	1.11	0.51	E2	16	8
232.pdf	-0.58	0.76	1.39	1.11	E3	5	19
187.pdf	-0.71	0.70	1.06	0.51	E2	12	12
170.pdf	-0.94	0.65	1.12	0.58	E3	6	18
194.pdf	-0.95	0.70	0.65	0.26	E3	4	20
172.pdf	-0.96	0.66	1.00	0.59	E3	6	18
188.pdf	-1.00	0.71	0.81	0.43	E2	18	5
229.pdf	-1.05	0.68	0.71	0.30	E2	16	8
169.pdf	-1.16	0.64	0.67	0.32	E3	5	18
174.pdf	-1.37	1.07	0.69	0.12	E3	1	22
180.pdf	-1.51	0.75	0.45	0.17	E1	18	5
219.pdf	-1.52	0.77	0.78	0.29	E1	19	5
265.pdf	-1.52	0.74	0.71	0.28	E2	18	6
117.pdf	-1.60	0.71	1.53	0.92	A2	15	9
206.pdf	-1.62	1.01	0.95	0.28	E3	3	21
221.pdf	-1.85	0.68	0.48	0.21	E1	19	5
404.pdf	-1.86	0.81	0.79	0.28	E3	3	21

241.pdf	-1.93	0.69	0.64	0.27	E3	5	19
165.pdf	-2.01	0.69	0.51	0.20	E2	16	8
190.pdf	-2.22	0.73	1.12	0.50	E2	13	11
134.pdf	-2.49	0.77	1.14	0.40	A2	12	11
116.pdf	-2.54	0.66	0.51	0.24	A2	16	8
142.pdf	-2.62	0.67	0.88	0.41	E1	15	9
223.pdf	-2.75	0.64	1.09	0.66	E1	15	8
166.pdf	-2.87	0.93	0.73	0.21	E2	12	12
181.pdf	-3.11	0.68	0.74	0.32	E1	14	9
222.pdf	-3.52	0.64	0.74	0.39	E1	15	9
225.pdf	-3.55	0.72	0.90	0.49	E1	14	9
173.pdf	-3.69	1.91	0.10	0.01	E3	0	24
144.pdf	-4.06	0.78	0.86	0.85	E2	10	14
186.pdf	-4.25	0.76	0.37	0.14	E1	16	8
224.pdf	-4.51	0.70	0.36	0.17	E1	15	9
112.pdf	-4.52	0.81	0.39	0.13	A2	8	16
182.pdf	-5.43	0.77	0.79	0.41	E1	8	16
260.pdf	-5.79	0.58	0.85	0.54	E1	11	13
184.pdf	-5.96	0.61	0.64	0.35	E1	10	14
177.pdf	-6.07	0.72	0.58	0.23	E1	10	14
401.pdf	-6.19	0.68	0.78	0.44	E1	12	12
220.pdf	-6.35	0.71	0.68	0.25	E1	9	15
262.pdf	-6.37	0.64	0.69	0.31	E1	8	16
101.pdf	-6.39	0.56	1.05	0.74	A1	11	13
228.pdf	-6.53	0.58	0.94	0.56	E1	10	14
179.pdf	-6.66	0.60	0.63	0.34	E1	11	13
261.pdf	-6.82	0.68	1.14	0.54	E1	11	13
400.pdf	-6.87	0.60	1.11	0.63	E1	9	15
183.pdf	-6.92	0.75	1.01	0.90	E1	6	18
227.pdf	-7.26	0.60	0.88	0.53	E1	8	16

226.pdf	-7.40	0.59	0.84	0.47	E1	8	16
263.pdf	-7.44	0.69	0.59	0.24	E1	5	19
161.pdf	-7.56	0.69	0.49	0.21	E1	5	18
264.pdf	-7.74	0.64	1.24	0.76	E1	8	15
114.pdf	-7.94	0.69	0.65	0.28	A1	6	18
132.pdf	-7.95	1.11	2.25	0.53	A2	4	19
160.pdf	-8.20	0.71	1.22	0.59	E1	6	16
178.pdf	-8.24	0.61	0.97	0.66	E1	5	19
162.pdf	-8.70	0.76	0.47	0.18	E1	5	18
402.pdf	-8.75	1.17	0.28	0.05	E2	1	23
140.pdf	-9.20	0.73	0.45	0.17	E1	3	21
163.pdf	-9.47	0.74	1.08	0.48	E1	4	20
130.pdf	-9.57	0.77	0.86	0.41	A1	3	21
126.pdf	-10.97	1.88	0.06	0.01	A1	0	24
127.pdf	-11.50	1.89	0.07	0.01	A1	0	24
129.pdf	-12.08	1.87	0.05	0.01	A1	0	24

Appendix 6: Descriptive statistics for script quality measures by script group and level

E1/A1

Script group	N	Median	Mean	Std. Dev.	Range	Min	Max
CEFR bench.	6	-10.27	-9.74	2.22	5.69	-12.08	-6.39
SfL scripts	22	-6.13	-5.50	2.39	7.96	-9.47	-1.51
ESOL ex.	2	-5.91	-5.91	4.66	6.59	-9.20	-2.62

E2/A2

Script group	N	Median	Mean	Std. Dev.	Range	Min	Max
CEFR bench.	4	-2.54	-3.82	2.54	6.35	-7.95	-1.6
SfL scripts	10	-0.64	-0.21	1.90	6.39	-2.22	4.17
ESOL ex.	2	-1.38	-1.38	3.79	5.37	-4.06	1.31

E3/B1

Script group	N	Median	Mean	Std. Dev.	Range	Min	Max
CEFR bench.	10	2.76	3.49	2.77	9.25	0.29	9.54
SfL scripts	24	0.84	1.21	2.26	7.73	-1.86	5.87
ESOL ex.	3	2.69	3.73	2.42	4.51	1.99	6.50

L1/B2

Script group	N	Median	Mean	Std. Dev.	Range	Min	Max
CEFR bench.	8	5.82	5.96	2.30	7.12	2.82	9.93
SfL scripts	18	4.85	4.38	2.04	7.08	0.85	7.93
ESOL ex.	3	8.68	7.85	1.51	2.65	6.11	8.76

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344

public.enquiries@ofqual.gov.uk

www.gov.uk/ofqual