RESEARCH AND ANALYSIS

Evaluating the impact of the introduction of reformed GCSE MFL assessments in 2018

Tim Stratton and Nadir Zanini

ofqual

Contents

1 Executive summary

Background and motivation

GCSE modern foreign languages (MFL) qualifications have recently been reformed. New French, German and Spanish specifications were introduced for first assessment in 2018. A number of changes have been made to GCSE MFL assessments as part of the reforms. The Department for Education (DfE) stipulated that reformed GCSE qualifications should have more demanding content. Ofqual also introduced changes to the structure of the assessments through regulation and guidance to exam boards, with the aim that they would provide a fairer representation of students' knowledge and skills in MFL.

Prior to the reformed assessments being taken, concerns were raised that some of the changes introduced would make the reformed qualifications overly difficult. In particular, the requirement that audio tracks for the listening assessment would include 'standard speech at near normal speed' raised concerns over a potential increase in speech speed from the previous assessments and therefore increased difficulty. Alongside this were concerns that there would be less time for students to formulate their answers. There was also a concern that the introduction of questions written in the target language may make these questions inaccessible to some students, potentially disadvantaging them. In addition to these changes to the individual assessments, qualification-level changes implemented included a reduction of non-exam assessment. Ofqual therefore committed to carrying out a technical evaluation of the reformed specifications to consider the impact of these changes, and whether there was any evidence that students had been disadvantaged by the changes (Jadhav, 2018).

The first aim of this research was to explore the impact that the specific changes introduced to MFL assessments in 2018 might have had on the difficulty of individual questions and assessment components. The study focusses on the key concerns raised by stakeholders about the reformed assessments, and whether there is any evidence that students taking these assessments in 2018 have been disadvantaged by the changes. The second aim was to establish whether the new assessments are functioning effectively and have improved with respect to classification accuracy (i.e. differentiation of students) at the component and qualification level.

Methodology

The main analysis was based on the comparison of the difficulty of assessments taken in 2017 and 2018. This analysis was performed at 'item' level ie the finest granularity of data available, in most cases this meant individual sub-questions. For each item in the listening and reading papers, facility (an indicator of item difficulty) and discrimination (how well individual items differentiate between students) were

computed. For discrimination scores there was little change between years and therefore the analysis focussed on facility scores.

Key features of concern were coded for each item, namely speed of speech, pause length and whether the question was written in the target language. Other item features which potentially affect students' performance on exam items were identified from an extensive search of the language testing literature. Some of these aspects were coded from the exam materials, others were rated by subject experts. A statistical model was used to identify which item features had an effect on difficulty. This was then compared against the item features that had substantially changed between 2017 and 2018.

It should be noted that this approach produces evidence on the relationship between certain item features and difficulty of assessment. It does not allow us to address the concerns raised by some stakeholders as to whether, as an example, the use of vocabulary in the assessment is appropriate. Further research may be needed to look at the validity of certain item features such as vocabulary use.

Summary of findings

As for the specific concerns raised by stakeholders, analysis showed that:

- The mean facility scores have generally decreased in 2018 suggesting an overall increase in difficulty. However, this increase is likely not due to the key features which were initially of concern.
- Speed of speech and pause length in the listening assessments had little effect on item difficulty and did not change substantially between years.
- The introduction of questions in the target language only had a significant impact on French reading assessments, but not to a degree where questions would likely become inaccessible.
- The increase in difficulty in 2018 appeared to be primarily due to an increase in the demand of the vocabulary used in the reading and listening texts and questions requiring more 'work' from students to answer the question (eg not being able to rely on spotting key words or phrases).
- The introduction of literary extract based questions, translation questions and the use of more short answer questions is likely to have also increased difficulty. These changes are in line with the intentional increase in the demand of content stipulated by DfE as part of the reforms to GCSEs.

The findings of this study suggest that the new assessments in 2018 are functioning effectively. By this we mean that the assessment is at an appropriate level of difficulty and is successful at differentiating students across a range of ability. Although, on average, students are obtaining fewer marks in the assessments, analysis indicated that grade boundaries had become more spread out in 2018, allowing better differentiation of students. For a few of the exam papers, the facility

scores and grade distributions from the 2018 assessments were quite low (potentially suggesting they were too difficult), but this is possibly due to the lack of familiarity with the new assessments.

As for the component-level analysis, this study suggests that:

- The changes to the writing assessment have improved the balance between the assessments, in terms of the weighting across assessments and the distribution of marks.
- Students are generally showing lower levels of attainment in the writing assessment since it has moved from controlled assessment to exam-based assessment. However, this has been balanced by an increase in attainment in both the reading and listening components, resulting in stable qualification-level outcomes.
- Due to better assessment functioning, students' probability of obtaining a C/4 (or above) or an A/7 (or above) is more similar between components in 2018, correlations between component marks is generally higher and each component is generally a better predictor of GCSE level outcomes, suggesting that GCSE grades will better reflect students' ability across the skills assessed.

Overall conclusion

Overall, from a technical functioning perspective, the new reformed assessments are functioning better than the pre-reform assessment. Despite this being a necessary criteria for a valid assessment, further research was undertaken by Ofqual to address whether these qualifications are valid in relation to their specific purpose (Ofqual, 2019). This report, however, shows that, with respect to the previous cohort, there is no evidence that students taking the GCSE MFL assessments in 2018 were disadvantaged by the changes introduced with the reform. In fact the reformed assessments are likely to be more reliable in classifying students by ability and produce a fairer representation of students' knowledge and skills.

2 Introduction

Modern foreign languages (MFL) were included in the programme of reforms to GCSEs implemented between 2015 and 2019. As a result, new French, German and Spanish specifications were introduced for first assessment in 2018[1]. Previously some issues had been raised with the functioning of the legacy assessments. The assessments introduced in 2018 were designed to overcome these issues so that the grade achieved by candidates would provide a fairer indication of students' preparedness. Some stakeholders, however, expressed concerns that the changes to the assessments would make the assessments more difficult, disadvantaging students taking the reformed assessments.

Once results from the first awards became available, Ofqual undertook research to evaluate the impact of changes to GCSEs in MFL on grade standards. The overall aim was to understand whether the assessments were fair to students taking MFL in 2018 and to ensure that they have not been disadvantaged by the changes to the assessments due to the reform.

Although some preliminary findings have already been publicly shared (Stratton, 2019), this report presents and discusses in detail the findings from this research. Before doing so, however, it is necessary to describe the main changes to the GCSE MFL assessments and to provide an overview of the context within which these changes have been implemented.

2.1 Changes to GCSE MFL

In addition to the introduction of a new 9 to 1 grade scale to allow greater differentiation in student ability, a series of changes have been made to GCSE MFL assessments as part of the reforms, both to the content and the structure of the assessments (summarised in Table 1). The Department for Education (DfE) stipulated that reformed GCSE qualifications have more demanding content to add stretch and challenge. Adjustments to the structure of the assessments were also implemented with the aim to improve the validity and reliability of the assessments.

[1] For other MFL subjects new specs will be available for first assessment in 2019 and 2020.

Table 1. *Summary of relevant changes to GCSE MFL listening and reading components.*

<table>
<tr><th></th><th>2017</th><th>2018</th></tr>
<tr><td>Weighting of non-exam assessment</td><td>- 40% exam
- 60% controlled assessment (Speaking and Writing)</td><td>- 75% exam
- 25% non-exam assessment (Speaking)</td></tr>
<tr><td>Tiering</td><td>Listening and reading are tiered at either Foundation Tier or Higher Tier; students can enter different tiers for listening and reading.
Speaking and writing are untiered.</td><td>Question papers and speaking assessments set at either Foundation Tier or Higher Tier.
No mixed tier entry permitted.</td></tr>
<tr><td>Length of listening assessments</td><td>No rules set.
Exam board approach was:<table><tr><th></th><th>AQA</th><th>Pearson</th><th>WJEC</th></tr><tr><td>FT</td><td>30</td><td>25</td><td>35</td></tr><tr><td>HT</td><td>40</td><td>35</td><td>45</td></tr><tr><td>For reading</td><td>+5</td><td>+5</td><td>+5</td></tr></table>Note: timing in minutes</td><td>- Foundation Tier 35 mins
- Higher Tier 45 mins
- For reading +5 mins</td></tr>
<tr><td>Listening</td><td>Listen and respond to different types of spoken language.</td><td>Listen to and understand clearly articulated, standard speech at near normal speed.</td></tr>
<tr><td>Reading</td><td>Read and respond to different types of written language.</td><td>To include authentic material and literary texts including… poems, letters, short stories, essays, novels or plays from contemporary and historical sources.
Translate a short passage from the assessed language into English.</td></tr>
<tr><td>Questions in assessed language</td><td>A minimum core vocabulary must include, where applicable, key words and phrases used in rubrics in the language.</td><td>Questions may be set in the assessed language or English, as appropriate to the task.
Questions should be set in the language in which the student is expected to respond.</td></tr>
</table>

In line with DfE content requirements, the reformed reading papers contain more authentic stimulus material including extracts from literary texts and short translation exercises. In the listening components students must "listen to and understand clearly articulated, standard speech at near normal speed" (DfE, 2015). Across all of the new components, students will also have to answer questions written in the target language, whereas they were previously all written in English. DfE requirements were included in Ofqual regulation (Ofqual, 2017), which states:

> *In listening (AO1) 20 - 30% of the marks must be awarded for responses to questions set in the assessed language.*
>
> *In reading (AO3) 30 - 40% of the marks must be awarded for responses to questions set in the assessed language.*
>
> *In writing (AO4) students will be required to express themselves solely in the assessed language. Questions may be asked in English where translation into the assessed language is required or where the context of the questions is detailed or complex.*

In addition to DfE requirements, a number of structural changes were introduced by Ofqual to improve the functioning of the assessments. The new MFL assessments have a reduced amount of non-exam assessment. In the previous specifications, 60% was non-exam assessment, covering the speaking and writing elements. In the new specifications only speaking is not assessed by written exam (instead it is an oral assessment carried out by teachers but marked externally) and the weighting of this is reduced to 25% of the overall MFL grade. Controlled assessment was reduced because, particularly in MFL, research indicated it had a detrimental effect on teaching and learning. Teachers of MFL indicated that the writing assessment was a test of memory skills rather than being a valid assessment of language skills (Ofqual, 2013).

Research also highlighted that in MFL controlled assessment made up a large proportion of the marks (Ofqual, 2013). In the reformed assessments the weighting of the components has been adjusted to give equal weighting to all components, examined and non-examined. Prior to reform, the four assessment components (reading, listening, writing, speaking) were weighted 20%, 20%, 30%, 30% respectively. In the reformed specification, the weighting of the speaking and writing assessments has been slightly reduced and each element is now equally weighted (25% each). The tiering structure has also been adjusted, as previously only the listening and reading components were tiered, whereas in the new specification all the components are tiered. In addition, students have to take all components in the same tier, whereas previously they were able to 'mix and match', although in practice only a few did.

2.2 Concerns around the difficulty of MFL

There have historically been concerns in the MFL community that assessments in these subjects, both at GCSE and A level, have 'severe grading' (eg Guardian, 2015; Guardian, 2019; TES, 2019). These concerns are related to the low uptake of MFL subjects in relation to other GCSEs and the negative trend in entries to French and German over the last few years. Since it became non-compulsory for students to take a foreign language at GCSE level in September 2004, there has been a steady decline in entry (Tinsley & Doležal, 2018, Churchward, 2019).[2] There is also concern that the perception of MFL subjects being severely graded is having a knock-on effect on take up at A level. However, it is likely that additional factors contribute to the low uptake of languages (Board & Tinsley, 2016; Tinsley & Doležal, 2018).

In 2017, Ofqual conducted research on the impact that the presence of native non-English speakers taking A levels in their own language had on MFL grading standards (Taylor & Zanini, 2017). The research led to an adjustment to grading standards in 2017 such that approximately 1% more students achieved a grade A or above. Ofqual has also previously published a tranche of work at A level exploring inter-subject comparability, which suggested there was not a compelling case for adjusting the A level standard, but did result in a one sided reporting tolerance to exam boards, essentially preventing the assessments become more difficult in future (Ofqual, 2018). At the time when this report is written, more work is being conducted by Ofqual to gather evidence to inform a decision on whether GCSE MFL standards should be adjusted.

Prior to the new GCSE assessments being taken, therefore, concerns were raised that some of the changes to the assessments would make the qualifications overly difficult. In particular, the stipulation in the new specification that students should be able to 'listen to and understand clearly articulated, standard speech at near normal speed' has raised concern over a potential increase in speech speed from the previous listening assessments and therefore increased difficulty. Alongside this are concerns that, due to the new regulation around the length of the listening assessments, there will be less time for students to formulate their answers. The other major concern raised was regarding the introduction of questions written in the target language. Previously all questions had been written in English, and so concerns were raised that, particularly for foundation students, this may make these 'target language questions' inaccessible, potentially disadvantaging students.

Ofqual have therefore committed to carrying out a technical evaluation of the reformed specifications to ensure that they were functioning adequately and that the material was accessible to students (Jadhav, 2018).

[2] The decline in entry to MFL had actually started before it became non-compulsory to take a language at GCSE. The decline, however, became more pronounced after 2004.

2.3 The impact of reform on student achievement

With any reform to qualifications, in the first year of assessment a small drop in performance (in terms of the number of marks achieved on the exam) is likely. This is due to teachers being less familiar with the nature and requirements of the new assessments, irrespective of any potential changes in demand of the assessment. This has been termed the 'Sawtooth effect'. Previous research suggests that it takes approximately 3 years for performance to return to previous levels after a change to assessments (Ofqual, 2016; Cuff et al., 2019).

In GCSEs (including MFL) and A levels, exam boards use predictions to maintain qualification standards over time and between boards in a subject. When entries are large enough, this approach uses predictions based on students' prior attainment at cohort level, so that any year-on-year change in the difficulty of assessments does not affect students' chances to achieve a certain grade (Taylor & Opposs, 2018). This means that, where the prior attainment of the cohort is stable, it is likely that a similar proportion of students will achieve each grade, compared to previous years.

In the first years of the reformed GCSEs, these predictions were used to carry forward the standards from the legacy GCSEs, so that students taking the reformed GCSEs were not disadvantaged with respect to those who took the qualification in 2017. Any sawtooth effect was likely to mean that students would perform slightly less well in the new assessments. However, it is still important to ensure that those assessments are functioning effectively. This includes ensuring that they are not systematically overly difficult (or easy), that they allow differentiation of students and ensuring that the assessments are a valid reflection of students' ability.

2.4 What do we mean by demand, difficulty and performance?

Key to this study is an understanding of what we mean by item difficulty, how it can be measured and what it tells us about an assessment. So here we define what in this report is meant by demand, difficulty and performance and how these features interrelate.

Generally, by *demand* we refer to an objective view of the complexity or comprehensibility of the assessment task irrespective of the students taking it. In the exam assessments considered here, the assessment task includes both stimulus material (text in the reading assessment and audio tracks in the listening) and the exam question. Features of both the question and stimulus material, and potentially the interaction between the two, can lead to differences in question demand. Adjustments to task demand are usually intentional, relating to features of the content and curriculum which students would be expected to know. An increase in demand in this case could be caused by using more complex or less familiar

vocabulary in the stimulus material, or making items require more work from the students by making answers to the questions less obvious. Demand may also be affected by features such as the style or type of question being asked, among other factors. An increase in assessment demand was part of the intention of the GCSE reforms. Increasing demand can improve the functioning of an assessment if it allows the more able students to show their knowledge and skills and therefore provides a greater spread of marks. This helps improve classification accuracy (ie students being correctly rewarded with the grade they deserve) through the spreading out of grade boundaries (Crocker & Algina, 2008).

Difficulty can be assessed by looking at how students collectively performed on an assessment or individual items. Sources of item difficulty (or easiness) are layered on top of item demands and may modify items to become easier or more difficult (Pollitt et al., 1985). They can be intentional (and valid), such as the command word used in an item or providing more or less scaffolding to a student. However, there can also be unintentional (and potentially invalid) sources of difficulty such as the wording of an item being confusing or if items require prior knowledge not relevant to the subject. If items become difficult for the wrong reasons (ie due to features not relevant to the intended scale of the assessment) or if items become inaccessible to some students preventing them from showing their ability, then this can cause an assessment to function poorly. Throughout this report we will use *facility* as a statistical index of item difficulty. In addition to the above, facility is related to the ability of the cohort taking the assessment, which needs to be taken into account when comparing assessments.

In this report we will also refer to *discrimination*, as the property of an item (or assessment) to differentiate between students of different underlying abilities. If the difficulty of an item (measured by its facility) represents the average performance of students, its discrimination gives an indication of how well the item distinguishes between students of different abilities. An assessment which is too easy or too difficult overall, where students on average receive a very high or very low proportion of the marks, is particularly problematic when it contains many items with low discrimination. This will result in grade boundaries becoming clumped together producing greater potential error in classifying students by ability.

Whereas demand, difficulty, facility and discrimination are all attributes of an item or of an assessment, *performance* refers to the quality of students' work. This can be quantified by the marks achieved by students on an assessment. Performance is strictly linked to difficulty (and therefore the demand) of the items. It also depends on students' ability and/or their preparation, which may be less effective immediately following reform due to both teachers' and students' lower familiarity with the exam structure and content.

2.5 Research aims

The overarching aim of this research is to evaluate the changes to the reformed MFL assessments and whether the assessments in 2018 are fair to students. More specifically, we will provide evidence on how the assessments are functioning and on how students taking the assessment in 2018 performed with respect to those taking the assessment in 2017. This research is divided into two strands of work.

The first strand explores the impact the specific changes introduced to MFL specifications in 2018 have had on the difficulty and discrimination of individual items and assessment components. Here we focus on the key concerns raised by stakeholders about the reformed assessments, namely: the impact of the potential change in the speed of the recordings in the listening assessment and the introduction of questions written in the target language in both the listening and reading assessments. However, this study will also consider other potential sources of difficulty in these assessments. It will evaluate whether they have changed due to reform and how they affected the difficulty and accessibility of the assessments in 2018.

The second strand evaluates the assessments at a component and assessment level. In order to evaluate how the structural changes to the assessment have impacted students we investigate how the relationships between students' performance on the different elements of the assessment have changed between 2017 and 2018 and whether this has differentially impacted students of different ability.

3 Methodology

3.1 Overview

The central analysis was based on the comparison of students' performance on assessments taken before and after the introduction of the reformed assessments.

The item level analysis carried out in subsequent sections focussed on the reading and listening assessments as these were examined components both pre and post reform. It was therefore possible to examine how individual features of theses assessments had changed with the reform and the impact this had on the assessments. The subsequent component level analysis included listening, reading and writing components to evaluate the relationship between the assessments and how this had been affected by the change in the writing assessment from controlled assessment to an exam. The speaking component was not considered in this study. Speaking is tested through a conversation, with some prompts. As such there are no clearly defined items that would have allowed a detailed analysis of exam functioning as was possible for the other components.

Exam boards provided item and student level data for each of the examined components from 2017 and 2018. Item level data on the mark each student obtained on each item of the assessment was provided for the listening and reading components. In this report we use 'item' to refer to the lowest level of question granularity for which data was available, in most cases this was at the sub-question level (eg 1a, 1b, 1ci, 1cii), and 'question' to refer to numbered questions (eg question 1, question 2) including all of the relevant sub-questions. Exam boards also provided data on outcomes at the component and qualification level, including grade boundaries. The analysis focused on 16-year-old students (calculated as age on 31st August in the year they took the exam) from England only.

With the data provided by exam boards it was possible to compute for each item in the listening and reading papers facility and discrimination scores. Facility and discrimination were then studied in relation to item features. These features include those relating to the key issues highlighted by stakeholders in advance of the new assessments (ie the language the question was written in, the speed of the speech in the listening tracks and the time left between tracks for students to write their answers) and other features identified by the literature that potentially affect students' performance. The analyses allowed identification of which item features were best at explaining the facility and discrimination of each item.

In the next section, a detailed description is given of how facility, discrimination and other item features are defined and computed, before describing the statistical analysis performed.

3.2 Item facility and discrimination

The data provided by exam boards was used to calculate facility scores, to be used as an index of item difficulty[3]. Facility scores take a value between 0 and 1 indicating the proportion of marks that all students obtained on that item out of the total number of marks available. For a one-mark item this simply translates to the proportion of students who correctly answered the item, for multi-mark items it gives an average score across students scaled between 0 and 1. A facility score of 1 therefore indicates that all students got the item completely right, whereas a facility score of 0 indicates that all students got the item completely wrong. Facility was calculated at the finest granularity that awarding organisations were able to provide data, which, in most cases, was at item level.

Although facility is a proxy for relative item difficulty, it can vary for a number of reasons. It is inherently related to the ability of the students taking the assessment, as more able students will be more likely to answer a particular item correctly, leading to a higher facility score for that item. In this study, prior attainment scores for each cohort were relatively stable within assessments between years allowing us to have confidence in comparing facility scores over time (see Appendix A). However, a measure of concurrent mean GCSE[4], as a proxy for average student ability, was included in the analysis to control for any differences in student ability. Facility may also change based on students' preparedness and familiarity with the exam. Given that 2018 was the first year of a new set of assessments, we may expect student performance to drop slightly and therefore we might expect to see lower facility scores in 2018. This will be explored in the analyses below.

Discrimination gives an indication of how well each item distinguishes between students of differing ability. A discrimination index was calculated as the correlation between students' scores on each item and their score on the overall test after removing the item in question. Discrimination scores can take a value between -1 and +1. Any item with a score lower than 0 suggests a very poorly functioning item as students who get the item right are predicted to get a lower score on the overall test. Generally scores range from 0 (a very poor predictor of overall performance) to 1 (a very strong predictor of overall performance). Discrimination is inherently linked to facility, as very hard or very easy items (with a facility near 0 or 1), are unlikely to discriminate between students. Identifying which question features are linked to high discrimination scores may aid to improve future assessments. Although the focus of

[3] Here we used a Classical Test Theory approach rather than Item Response Theory due to the lack of linking items/students between exam papers and due to the relative ease in calculation and interpretation of facility scores. When Rasch measures of item difficulty were calculated for individual assessments, they were highly correlated with facility scores.

[4] This was calculated for each student by converting their GCSE grades, taken in the same year as their MFL grade, to a numeric scale and taking the mean, then for each assessment taking the mean of that score for the cohort taking the assessment.

this report is the impact of changes between 2017 and 2018 on item difficulty, discrimination is also evaluated as it is key to the functioning of assessments.

3.3 Features affecting difficulty

Item features which potentially affect the difficulty of exam items were identified from an extensive search of the language testing literature (Crisp and Sweiry, 2005; Pollitt et al., 2007; Ahmed and Pollitt, 1999; Fisher-Hoch et al., 1997; Laufer & Nation, 1995; Ure, 1971; Bloomfield et al., 2010; Rupp et al, 2001; Pollitt et al., 1985; El Masri et al., 2017). Where possible, item features were coded by the project researchers from the exam materials, such as whether a picture was included with the item, number of words in the item prompt and the question topic. In other cases features were scored by experts in the target language using their judgement on the basis of their knowledge of GCSE MFL specifications and experience of how difficult 16 year old students find questions. This represents an attempt to overcome limitations highlighted in previous research on the item features affecting students' performance in two ways.

Firstly, in many previous studies (among the most recent ones, see El Masri et al., 2017) the objective coding of linguistic features was mainly considered. This failed to account for much variance in item difficulty and often resulted in complex models with a very high number of variables with complex interactions. This makes the analysis difficult to interpret and limits its use to improve assessments. In the current study, a combination of objective measures and subject expert judgement were used to help unpack the factors affecting the difficulty of items in MFL assessments while still being usefully interpretable. The intention was to identify subtler language features, which may be more subjective to judge, in an attempt to account for more of the variance in item difficulty.

Second, in using subject experts to score specific aspects of the items, we also asked them to make a more holistic consideration of the items, in order to make sure they accounted for additional aspects of difficulty. The command words used, the nature of the task and how these aspects interact with the target students are aspects that cannot be captured by objective coding and require an element of subjectivity. Controlling for these features then allows us to more clearly identify the impact on difficulty (and not only on demand) of the features which have changed due to reform.

3.3.1 Subject expert scores

Three subject experts were used for each language to provide expert ratings of features of the individual items that could potentially affect difficulty. Experts had experience of both teaching and assessing the target language at GCSE level.

Experts were asked to score each item on a series of 1-5 scales. Scales included features that the literature suggests can impact the difficulty of a language item (Lumley et al., 2012; Carr, 2006; Wauters et al., 2011; Pollitt et al., 1998), which could be reliably scored but required a degree of judgement. These scales are shown in Table 2.

For each scale a lower score was hypothesised to indicate a feature of easier items and a higher score more difficult items. Scales were refined and subject experts were standardised on the scoring system at a one day meeting, following which scoring was carried out by the experts at home. When scoring each item, subject experts were asked to consider all the relevant stimulus material and the text a student would need to read to answer that item. For listening items, in addition to the exam paper, the subject experts considered the audio and transcript of each track. For these assessments it was not possible to consider the demand of the stimulus material and associated question text separately. It is necessary to consider the whole task which includes the relevant parts of the stimulus text, intentionally distracting parts of the stimulus text and relevant text included in the question/answer section of the exam paper. Experts were asked to consider each item and each scale separately to ensure scores were independent of one another.

Table 2. *Summary scales scored by expert judges for each exam item and intra-class correlation (ICC) coefficients of judges scores, by language.*

Scale No.	Prompt	ICC		
		French	German	Spanish
S1	Score the overall difficulty of the vocabulary from 1 (easy) to 5 (very hard)	0.82	0.85	0.74
S2	Score how familiar students are likely to be with the vocabulary used in the question from 1 (very familiar) to 5 (very unfamiliar)	0.85	0.76	0.76
S3	Score the difficulty of the grammar (sentence structure/syntax/tenses) from 1 (simple) to 5 (very complex)	0.85	0.86	0.70
S4	Score the likelihood that students will be familiar with the topic of this question from 1 (very familiar) to 5 (very unfamiliar)	0.85	0.84	0.79
S5	Score how concrete or abstract the subject is from 1 (very concrete, eg objects/places) to 5 (very abstract, eg thoughts/emotions/ideas)	0.85	0.85	0.74
S6	Score how difficult is it to extract the information required to answer the question from 1 (all information required is easy to pick out/locate) to 5 (information is very diffuse or needs to be interpreted to respond)	0.78	0.84	0.64

Before utilising the scores in any further statistical analysis, inter-rater reliability was checked using intra-class correlation coefficients (ICC). An ICC score between 0.6 and 0.74 is usually considered good, and over 0.75 excellent (Cicchetti, 1994). In all cases scores were above 0.6, and the majority over 0.75 indicating good consistency between the judges in the rating of items, which confirms the features they were scoring were adequately defined. Once it was confirmed the ratings were reliable between judges, an average was taken as the score for the item for use in further analysis.

3.3.2 Other item features

Additional features of the items which may affect difficulty and could be reasonably objectively identified were coded by the research team. These include: whether a picture is used, topic, item type and instruction language. See Table 3 for a detailed description of each of them.

Table 3. *Features of written items scored.*

Feature	**Description**
Picture	Is a picture included with the item? (Y/N)
Topic	Broad topic area of the item. Topic areas used were based on exam board specifications, but for consistency across exam boards they were condensed to: Holidays, Home and Environment, Leisure, Lifestyle, Work and Education and Literary Extract (including items based on a literary extract; reading paper only).
Item type	Type of item. Reduced to: - Multiple Choice Question (requiring selecting a single correct right answer) - Multiple Selection (requiring selection of multiple correct responses/images from a list) - Blanks (requiring selecting the right word/phrase(s) to complete a sentence/passage - Short answer (requiring a written answer) - Matching (requiring matching a selection of words/statements/pictures) - Names (requiring matching a statement/image to a name) - Translation (requiring translating a short passage from the target language to English; reading paper only).
Instruction language	Language the item is written in. English or target language (French, German, Spanish).

For the listening components, the additional variables in Table 4 were extracted from each audio track. Scores for the gender(s) of speakers, track length, pause length and time between tracks relate to the total audio for each item.

Table 4. *Features of listening item audio scored.*

Feature	Description
Gender of speaker(s)	All male, all female or both genders.
Track length	Total time of stimulus audio in seconds.
Pause length	Time between repeats of audio track in seconds.
Time between tracks	Time between the end of the audio track and the start of the next track in seconds.

Lexical features were also calculated from the target language text for each item. For the listening items, this included the text from the audio transcript, and for the reading items, this was all the text which needed to be read to answer each item. Lexical features coded are detailed in Table 5. Any text in the target language which would need to be read to answer the item was included in the analysis, excluding instruction text. In some cases this meant the same text was reused for multiple items or sub-items.

Table 5. *Lexical features of text used in items scored.*

Feature	Description
Word count	Total number of words.
Sentence Count	Number of sentences (ending in a full stop, question mark or exclamation mark). Titles were considered as individual sentences. *Used to calculate words per sentence but not included in analyses.*
Words per sentence	Mean number of words per sentence.
Lexical Variety	Proportion of words which are unique within the text.
Lexical Density	Proportion of words which are 'content words', ie nouns, adjectives, verbs and adverbs.
Lexical unfamiliarity	Proportion of words which are taken directly from the vocabulary list in the specification. Reverse coded so a higher score indicates more unfamiliar words.
Words per second	For audio tracks only. Number of words spoken per second.

3.4 Analysis techniques

In this report we use a combination of descriptive statistics to look at the frequencies and averages of various features, bi-variate tests of difference, and regression analysis. The latter allows us to explore the relationship between the identified

factors and difficulty/discrimination. A technical description of the regression analysis used is given in the next section, followed by how this is applied to evaluate which factors affect item facility and discrimination.

3.4.1 Multivariate regression analysis

The use of multivariate regression analysis allows us to study the link between a dependant variable, y (for example, facility score of an item), and a number (k) of independent variables, say $x_1, x_2, \ldots, x_k$ (for example, the language of the question, word count or item type). The great advantage provided by the approach taken is that it allows us to draw conclusions on the *marginal effect* of x_1 on y, that is the impact of a unit change in x_1 on y, once the other factors $x_2, \ldots, x_k$ are controlled for. In other words, this provides information on the relationship between y and x_1 once the other factors $x_2, \ldots, x_k$ are held fixed.

We use different types of regression models. All those used in this report take the form:

$$y_i = F(\beta_0 + \beta_1 x_1 + \beta_1 x_1 + \ldots + \beta_k x_k)_i + u_i$$

The subscript i indicates each item and u is an error term. F is a probability function needed to take into account the distribution of the dependant variable y.

In the case of item facility, y only assumes values in the range 0-1, in which case a beta function is used for F. Beta regression allows the estimation of probabilities or proportions between 0 and 1, while allowing some variability in the distribution of the data. In the case of analysing the probability of a student achieving a certain grade (C/4 and above, or A/7 and above), the dependent variable may assume only the value 0 or 1 (achieved or not), in this case a binomial logistic regression is used instead. Where we have looked at other continuous variables as the dependent variable (speed of speech, pause length or discrimination) a simple linear regression model is used.

If all the variables affecting y are included in the regression model then β (or its transformation according to F) yields the unbiased estimate of the marginal effect of each x on y, once the other factors are controlled for. As it is impossible to ensure that all variables affecting y are observable and included in the regression model, the estimate of β is interpreted as the measure of association between each x and y, net of the effect of the other factors included in the model specification.

For linear models (those where the Identity function is used for F because a transformation is not needed), we report the estimates of the β coefficients associated with each variable. These coefficients indicate, after controlling for all other variables in the model, how much change, on average, we expect in the dependent variable y for each unit change in the relevant independent variable x. Positive values of β indicate an increase in y for each unit increase in x, negative values a decrease in y for each unit increase in x.

For beta and logistic regression models we present coefficients as *odds ratios* for ease of interpretation. Odds ratios indicate on average how much the dependant variable should be multiplied by for each unit increase in the independent variable. In this case values over 1 represent an increase in y for each unit increase in x, whereas values below 1 represent an expected decrease in y for each unit increase in x.

3.4.2 Predicting item facility and discrimination

Regression models were used primarily to identify which item features affected item facility (as a proxy for relative item difficulty). However, the same series of models were produced to evaluate whether we could identify which features affected the discrimination of items. The only change between these sets of models was whether facility or discrimination was included as the dependent variable and utilising beta regression models for facility and linear regression models for discrimination.

Due to high correlations between the different scales produced by the subject experts, only three of the six scales were included in the final modelling; *S1 – Vocabulary Difficulty, S3 – Grammar Difficulty and S6 – Difficulty to Extract Key Information* ('work' required by the student). Including highly correlated variables in a regression model causes multi-collinearity and subsequent difficulty in fitting the model and interpreting model outputs. Therefore, only these three scales which were not so highly correlated with each other ($r < 0.9$) were included, while retaining the scale which was most highly correlated with facility (S1). All other item features were included in the models as independent variables. The modelling procedure was performed as follows.

A regression model was fitted for each paper type (reading or listening) for each language (French, German or Spanish) at each tier (higher or foundation), totalling 12 separate models. A separate model was run for each tier as there are substantial differences in the prior attainment of students between tiers, but not between exam boards or years (see Appendix 8.1). The interaction between prior attainment, item features and facility may not be linear and so including tier as a covariate may not adequately represent this relationship. The assessments for each language (French, German, Spanish) include exam papers from three exam boards offering these assessments in 2017 and 2018 (AQA, Pearson, WJEC). The exam boards have been anonymised as EB-A, EB-B and EB-C in the results as the intention of this research is not to look at differences between exam boards. Hence throughout the analysis there may be some features which are more prevalent in some exam boards' assessments than others. In these cases the analysis will identify an average effect across exam boards, but further work would be required to investigate how these features have changed in specific assessments. However, it is reasonable to expect that these assessments from different exam boards are similar as all assessments are designed and accredited against the same criteria outlined by

Ofqual and qualification standards between exam boards are aligned by the use of predictions based on a national matrix.

Initially, data was modelled using just the component level information (awarding organisation and year) and the key features of interest (question language, speech speed, and pause length) and including mean GCSE score of all students in the assessment to control for any differences in student ability. This basic model was then compared to a full model including all of the other features described (see table 6). This approach allows us to observe if those key features are the main causes of variation in facility scores, or if the other features have a greater impact on item facility. Including additional features may also highlight that the initially observed effects of variables may change once the effect of other factors are accounted for. Furthermore, including year as a covariate allows us to capture factors (eg teaching quality) that might have changed over time and avoid attributing this effect to other features of the assessments. A similar argument can be used for the inclusion of exam boards.

Table 6. Details of variables included in the basic and full models of facility and discrimination.

	Reading		Listening	
	Basic Model	Full Model	Basic Model	Full Model
Year	✓	✓	✓	✓
Exam Board	✓	✓	✓	✓
Mean GCSE	✓	✓	✓	✓
Instruction language	✓	✓	✓	✓
Speed of speech			✓	✓
Pause length			✓	✓
Time between tracks			✓	✓
S1 (vocab. difficulty)		✓		✓
S3 (grammar difficulty)		✓		✓
S6 (work required)		✓		✓
Word count		✓		✓
Words per sentence		✓		✓
Lexical variety		✓		✓
Lexical density		✓		✓
Lexical unfamiliarity		✓		✓
Pictures		✓		✓
Topic		✓		✓
Item type		✓		✓
Track length				✓
Gender of speaker(s)				✓

3.4.3 Evaluating changes in item features in 2018

Our investigation was designed to assess whether the features which have an impact on item difficulty have changed with the reforms. If item features had an impact on item difficulty but didn't significantly change in frequency between 2017 assessments and 2018 then they are unlikely to have affected students taking assessments in 2018. Similarly, if features changed in their frequency or magnitude but had little impact on item difficulty then they are also unlikely to have had an impact on the assessment difficulty in 2018 compared with 2017.

For the key features of interest we include a series of linear regression models[5] to highlight how they have changed between years after controlling for board and tier, by including board, year and tier as independent variables. For each of the other item features, we evaluated if they had substantially changed with the reforms first by using descriptive statistics of means and standard deviations for quantitative variables or by using frequencies for categorical variables. We then used statistical tests to identify if the magnitude of the change is likely to be consequential. For quantitative variables we used a series of t-tests and for categorical variables we used proportion tests (a variant of chi-squared).

[5] Key figures from these models are referred to in text but details are not included in the appendices. Full model details are available upon request.

4 Results – Factors affecting item functioning

Initially, we present some descriptive statistics and charts of the key variables in the analysis. We begin with distributions of facility and discrimination scores to give an overview of the assessments, then we present the distributions of key item features of concern between years and components. Subsequently we investigate which features have changed between the 2017 and 2018 assessments. We then address how some of these variables interact, by looking at how facility and discrimination scores and omit rates (proportion of students not attempting each item) relate to question language. Finally we present the regression analyses allowing us to identify the net effect of each of the variables of interest on facility and discrimination while controlling for other potentially confounding effects. Further descriptive features of the assessments including number of items, mean GCSE scores and number of students can be seen in Appendix 8.1.

4.1 Assessment differences in facility and discrimination

4.1.1 Facility

A general indication of assessment difficulty can be visualised by looking at the distribution of item facility scores in each assessment (Figure 1). It is commonly considered that facility scores for most items on a test should fall within the range of 0.3 – 0.8. Outside of these bounds, items are less likely to discriminate usefully amongst the majority of the target students. A facility score under 0.3 suggests an item may be too difficult for the cohort taking the assessment, as students on average obtained less than 30% of the marks available on the item. Similarly, items with a score over 0.8 may be too easy as on average students obtained over 80% of the marks available on the item.

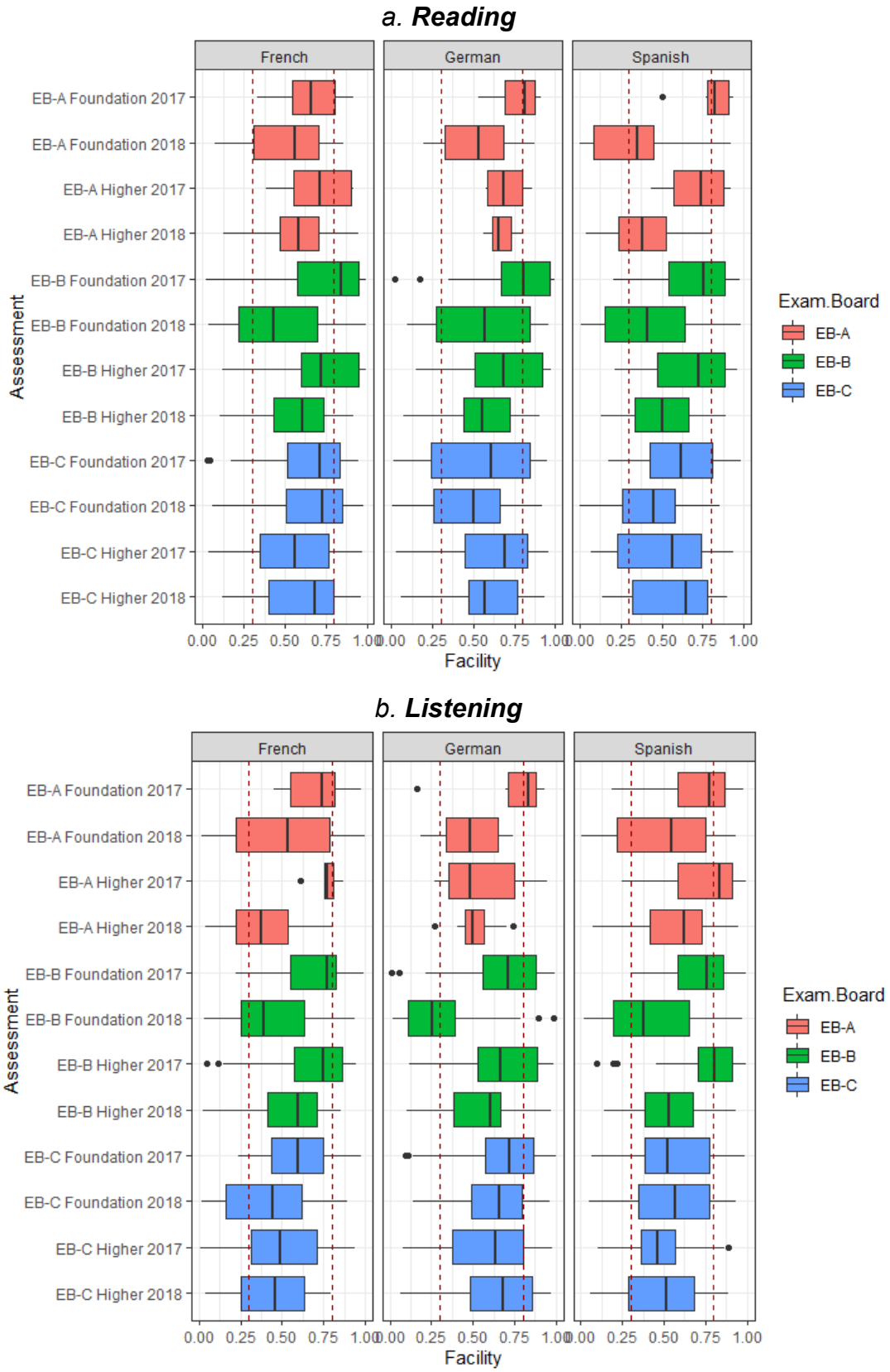

Figure 1. *Facility distribution of reading (a.) and listening (b.) assessments for each exam board and tier in 2017 and 2018.*

Note: Dotted red lines indicate lower (0.3) and upper (0.8) bounds of ideal facility scores.

Figure 1 shows that the majority of items across all assessments fell within the appropriate range, although there are a few points of concern. The majority of EB-B's 2017 papers and the EB-A foundation listening and reading papers for German and Spanish appear to be consistently too easy for the candidates taking the assessments. The corresponding 2018 papers have a much better distribution as on average item facility scores have decreased. However, a few papers in 2018 may have moved too far in the other direction and become overly challenging; EB-A foundation Spanish reading and EB-B foundation listening in German and Spanish. Given the potential sawtooth effect, this may improve in future years as students and teachers become more familiar with the reformed content meaning average performance increases slightly, without any change in the assessment difficulty.

After controlling for board and tier, a beta regression model indicated that overall, items had lower facility scores and were therefore more difficult in 2018. French reading assessments showed an mean decrease in facility by 31%, German by 43% and Spanish by 50%. For listening assessments, facility decreased by 40% for both French and German assessments and 48% for Spanish. All analyses significant at the $p<0.001$ level.

4.1.2 Discrimination

Discrimination scores give an indication of how well an item differentiates between students. Minimum acceptable discrimination is usually considered between 0.1 and 0.2 (Haladyna & Rodriguez, 2013), although even scores below this can be useful if they are consistently linked to student ability or if they are important to the construct or scale that the assessment aims to measure.

Figure 2 shows that almost all items across all assessments exceeded the minimum threshold of 0.1 and the vast majority had a discrimination over 0.2, suggesting that they were effectively helping to discriminate students. Generally there were no obvious shifts in discrimination scores between 2017 and 2018. Although for EB-A it appears as though there has been a general decrease in discrimination scores in 2018, this is likely due to the data in 2017 being at question rather than sub-question level for these assessments. This causes each individual mark on an item to be more strongly linked to total assessment mark and therefore having a higher discrimination score. For EB-C listening assessments there appears to be a slight increase in discrimination in 2018.

A linear model controlling for board and tier indicated that for reading assessments there had not been an overall significant change in discrimination. For listening, the picture was more mixed. For Spanish listening papers, discrimination had a slight, but statistically significant increase by a mean of 0.02 (SE=0.01, $p<0.05$). German listening papers also had an increase in discrimination by a mean of 0.05 (SE=0.01, $p<0.001$). For French, discrimination in the listening papers actually decreased by a mean of 0.03 (SE=0.01, $p<0.05$). This suggests the reforms have had little impact on the ability of the exams to differentiate between students at item level.

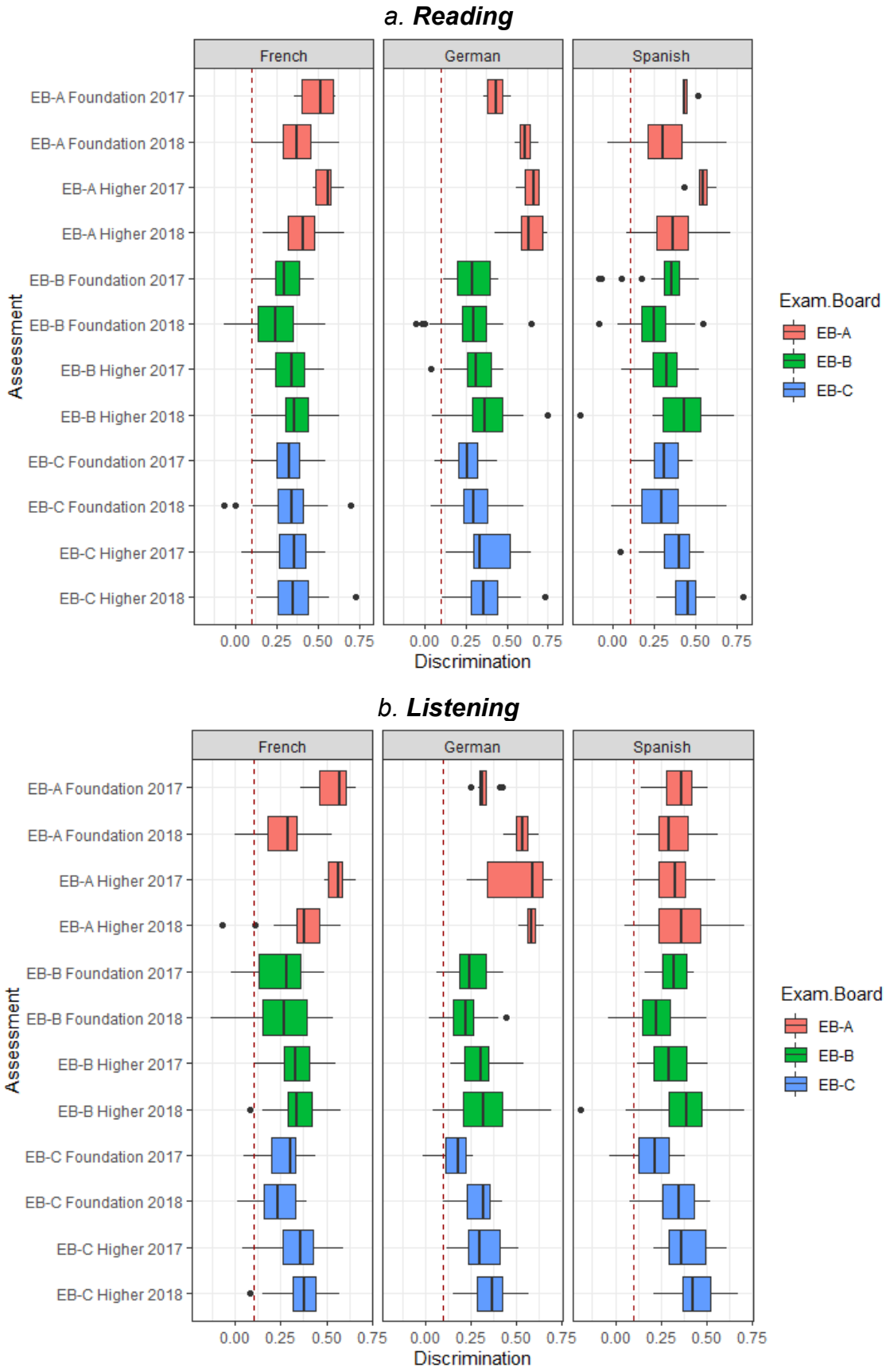

Figure 2. *Discrimination distribution of reading (a.) and listening (b.) assessments for each exam board and tier in 2017 and 2018.*

Note: Dotted red lines indicate lower (0.1) bound of ideal discrimination scores.

4.2 Differences in key item features

4.2.1 Speed of speech

One of the concerns with the new assessments was that, with the new requirements, the speed of speech would increase to a degree where students may struggle to understand and adequately respond to the items. The distribution of speech speed calculated as words per second, for each track, is shown visually for each of the assessments in Figure 3.

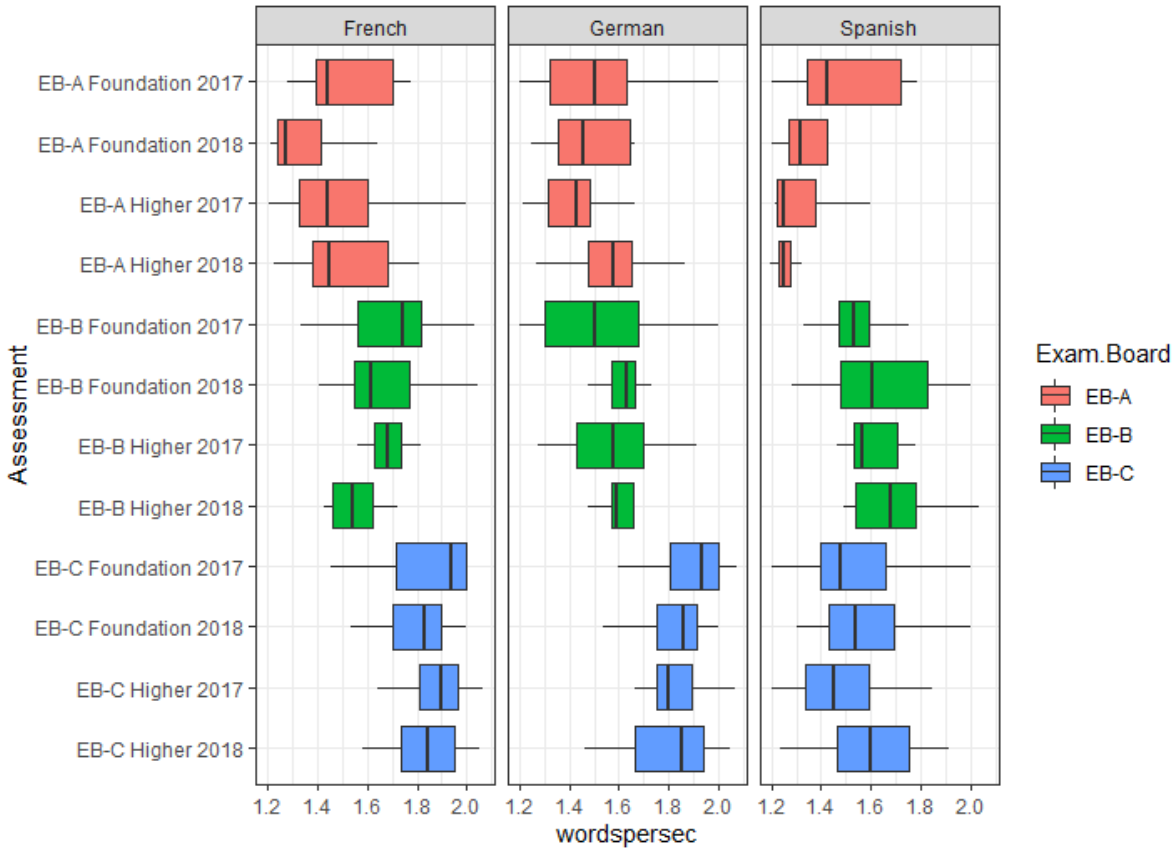

Figure 3. *Average number of words per second for each track in the listening assessments.*

Note: Outliers removed for clarity.

Speed of speech differed more across boards than within each board's assessments or between years. Speech speed in the EB-A assessments was generally the slowest and EB-C's the fastest. A linear regression model indicated that, after controlling for board and tier, on average speed of speech had decreased for the French assessments between 2017 and 2018 (with mean of 0.13 less words per second, SD=0.03, $p<0.001$) and this change was consistent across boards. However, for Spanish assessments the analysis indicated that the speech speed had a mean increase of 0.07 words per second in 2018 (SD=0.02, $p<0.01$). This increase

was not consistent between boards with EB-C increasing by 0.19 words per second, EB-B increasing by 0.04, but EB-As decreasing by 0.04 words per second. There was no significant difference in speech speed in German between years. The data therefore shows that overall there was no substantive change in speed of speech due to the reform.

4.2.2 Pause length

A further concern was that, given the potential change in speech speed and the stipulations regarding the length of the assessments, the time left for students to consider and write their answers would be reduced. This 'pause length' was considered in two ways. First, it was considered as the time between the repeats of each audio track. For all boards, each target language audio track was repeated twice and in this case pause length was considered as the time between the end of the first repeat and the start of the next. Second, it was considered as the time between the end of the second repeat of each audio track and the first repeat of the next audio track (time between tracks). Figure 4 presents the second option (the first showed a very similar pattern). Both time intervals were considered in the multivariate analysis.

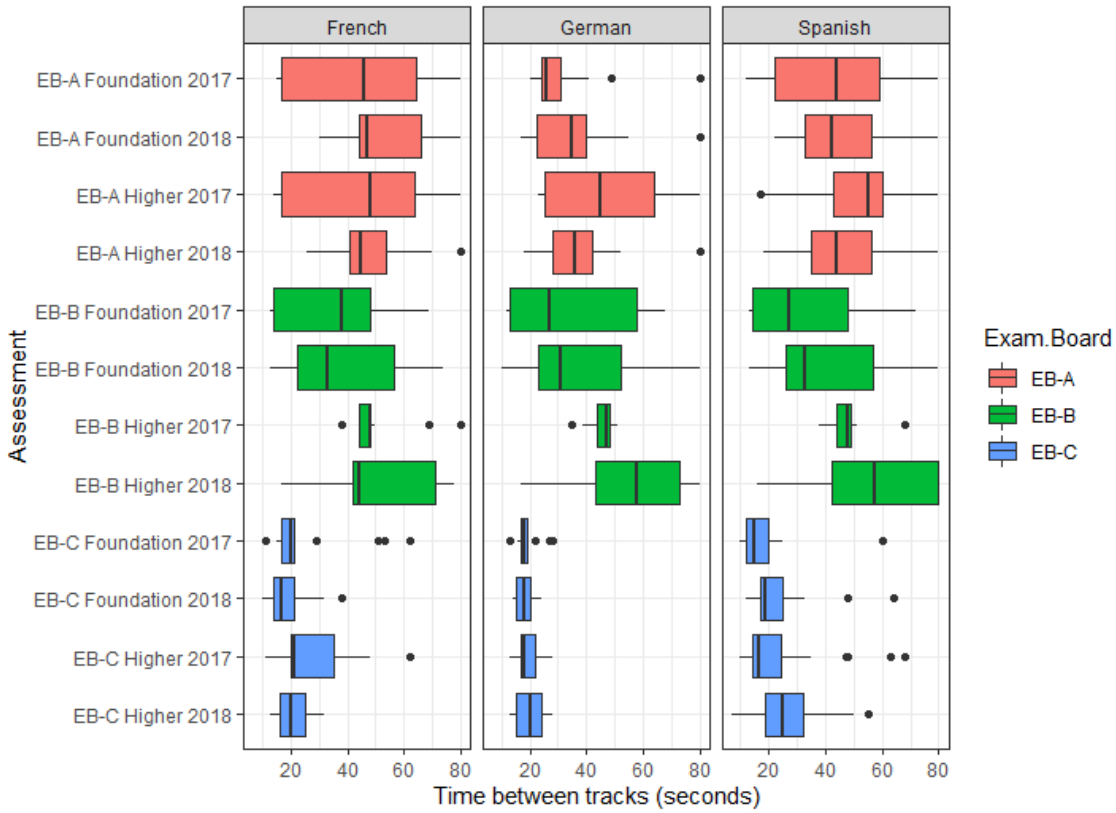

Figure 4. *Time left between the end of one spoken track and the start of the next for students to write answers for each track in the listening assessments.*

Mean pause length differed between exam boards, which may be related to the length of the audio tracks in each board. However, a regression model indicated that, after controlling for board and tier there had been no significant change in pause length (calculated as either time between tracks or time between repeats) between 2017 and 2018 for any language.

4.2.3 Target language questions

4.2.3.1 Omit rates

The accessibility of target language questions was first considered by looking at omit rates – the proportion of students who did not attempt each item. If students are not able to understand the question instructions in the target language then we may find an increase in students not attempting these items. Unfortunately omit rate data was not available for all assessments, but for cases where data was available, omit rates are shown in Figure 5.

For French and German reading assessments, a linear model indicated there was no significant difference in omit rates between items in English and items in the target language. For Spanish, a linear model predicted a significantly lower overall omit rate for items written in Spanish than English (-0.03, SD=0.01, $p<0.01$). This effect is the opposite of what might be expect if items in the target language are inaccessible to students and is likely due to the high omit rate of items written in English in the EB-A foundation paper.

For listening assessments, a linear model of omit rates by target language suggests that, after controlling for board and tier, omit rates are on average 3% higher (SD=0.01, $p<0.001$) where the question is written in French than English. For Spanish or German there was no significant difference in omit rates. This effect may be mainly due to the EB-C foundation French assessment which has a particularly high proportion of target language items not attempted. A further inspection of the data revealed that these were three of the last four items on the paper, two of which were also common items with the higher tier paper. These items also had low facility scores (only 2-10% of students got these items correct), which may suggest an issue. However, the common items will have been the most difficult questions on the foundation paper, targeted at grades 4 and 5.

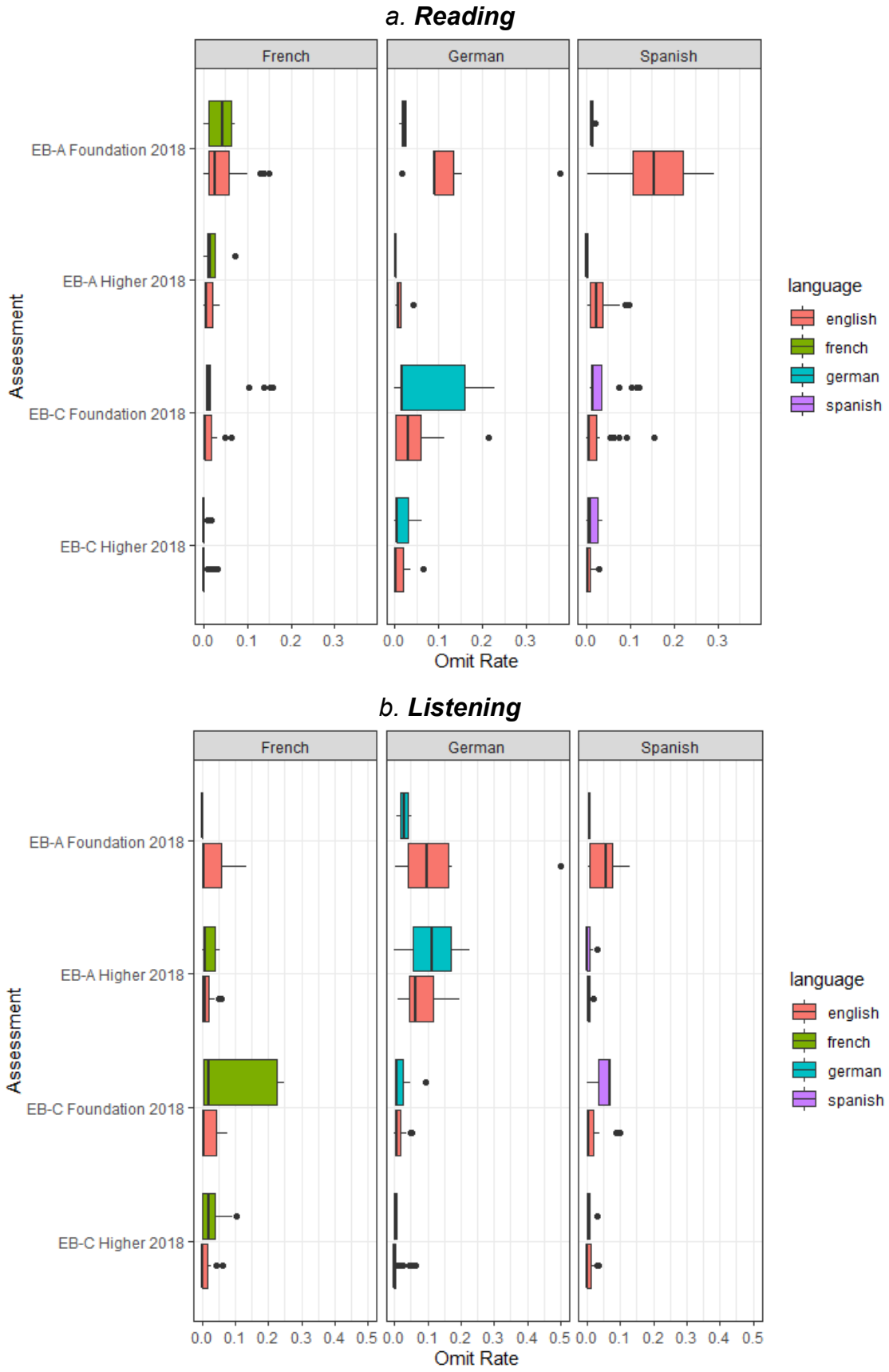

Figure 5. *Omit rates for each item in the EB-C and EB-A 2018 French reading assessments.*

4.2.3.2 Facility scores

Facility scores are generally considered a better indicator of item difficulty than omit rates as items can be omitted for a number of reasons, and as in this case omit data was not available for all assessments. The facility score distribution by language is presented for each assessment in Figure 6. If target language items are inherently more difficult than English items then we would expect lower facility scores for target language items than English items.

Visually, it can be seen that there is no clear pattern in the relationship between item language and item facility. In a number of cases, items in the target language appear to be those where students performed less well (such as EB-B foundation reading in all languages). In other cases the reverse appears to be true (EB-A foundation reading and listening papers in all languages). From these basic statistics, however, it is not possible to determine whether the students' poor performance in these items is due to the language used in the question or to other confounding factors which may include intentional differences in demand due to, for example, item type or the vocabulary used in the item text. As there were no stipulations over which questions should be in the target language, exam boards were free to use target language in combination with a number of other item features that might have impacted on the performance of students in addition to the language used. These complexities are addressed by the regression analysis, the results of which are presented in section 4.4.

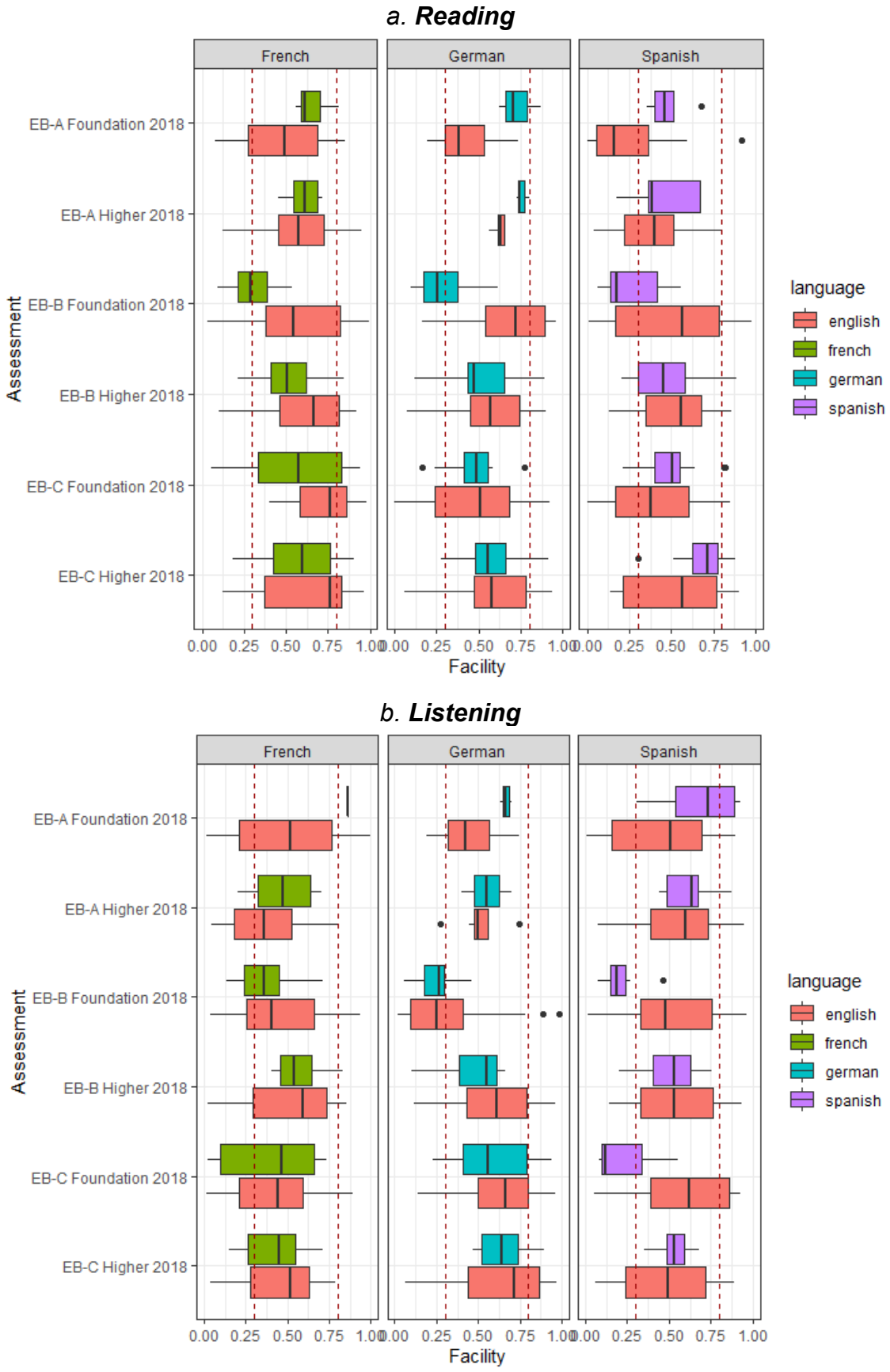

Figure 6. *Facility scores by item for all 2018 listening assessments split by question language*.

4.3 Other assessment changes between years

To identify if any other aspects of the assessments significantly changed between years, the mean and standard error of scores for each continuous variable of interest was calculated within each year and a t-test carried out to identify significant changes between years. For categorical variables the proportion of items in each category is given in each year and a proportion test (a variant of chi squared) was carried out for each subcategory to identify significant changes between years. For this analysis all exam boards' data was combined for each language and paper. Scores for French reading and listening assessments are shown below Table 7 and Table 8, results for Spanish and German can be found in Appendix B.

In 2018 across the reading and listening elements for all languages there has been a general increase in the rating provided by subject experts for vocabulary difficulty, grammar difficulty and the 'work required' for individual items, although for French this was mostly in the foundation papers and for German no significant change was seen in the higher reading paper. The increase in average work required is likely to be linked to the overall increase in the use of short answer written responses and a general decrease in the use of matching type questions. Foundation papers across languages and assessments had a general increase in average word count in 2018, which is also reflected in an increase in the average number of words per sentence in these assessments. Interestingly, given the increased ratings for vocabulary difficulty in 2018, there was also a general decrease in lexical variety in foundation assessments in 2018, meaning each item is using a smaller number of unique words. However, this would imply that they are, on average, more complex words.

Another notable change is that in almost every assessment there has been a decrease in the use of pictures included in questions in 2018. This may again be linked to the reduction in matching type questions which often include picture prompts.

Table 7. *Descriptive statistics of item variables in 2017 and 2018 French reading assessments.*

	Foundation			Higher		
Variable	Mean(SE) 2017	Mean(SE) 2018	T-test	Mean(SE) 2017	Mean(SE) 2018	T-test
Discrimination	0.32 (0.01)	0.31 (0.01)	-0.71	0.35 (0.01)	0.37 (0.01)	1.29
Facility	**0.69 (0.03)**	**0.55 (0.02)**	**-3.57 *****	0.64 (0.03)	0.59 (0.02)	-1.57
S1 (vocab. diff.)	**1.50 (0.06)**	**2.01 (0.06)**	**5.96 *****	2.56 (0.09)	2.72 (0.06)	1.63
S3 (grammar diff.)	**1.47 (0.06)**	**1.93 (0.06)**	**5.04 *****	2.58 (0.09)	2.72 (0.05)	1.42
S6 (work required)	**1.33 (0.06)**	**1.88 (0.06)**	**6.03 *****	**2.32 (0.09)**	**2.64 (0.06)**	**2.92 ****
No. Words	**58.16 (5.90)**	**84.22 (2.68)**	**4.53 *****	145.50 (9.35)	138.19 (3.62)	-0.84
Words per sentence	**8.59 (0.54)**	**10.66 (0.32)**	**3.49 *****	13.53 (0.77)	13.03 (0.24)	-0.73
Lexical variety	**0.83 (0.01)**	**0.75 (0.01)**	**-5.37 *****	**0.73 (0.01)**	**0.69 (0.01)**	**-3.44 *****
Lexical density	0.60 (0.01)	0.58 (0.00)	-1.69	**0.59 (0.01)**	**0.56 (0)**	**-3.55 *****
Lexical familiarity	**0.20 (0.02)**	**0.25 (0.01)**	**3.76 *****	0.25 (0.01)	0.27 (0.01)	1.81
	Proportion 2017	Proportion 2018	Chi-Squared	Proportion 2017	Proportion 2018	Chi-Squared
Instruction language - French	**0.00**	**0.32**	**28.27 *****	**0.00**	**0.33**	**29.66 *****
Picture included	0.17	0.07	3.56	0.13	0.10	0.20
Topic						
Holidays	0.07	0.03	0.54	0.09	0.03	0.20
Home and environment	0.33	0.20	3.19	0.29	0.24	2.01
Leisure	0.17	0.16	0.01	0.17	0.01	0.40
Lifestyle	**0.29**	**0.14**	**5.76 ***	**0.21**	**0.21**	0.00
Extract	**0.00**	**0.23**	**18.47** ***	0.00	0.21	**16.15 *****
Work and Education	0.14	0.24	1.97	**0.24**	**0.29**	**17.11 *****
Item Type						
Blanks	**0.00**	**0.08**	**4.96** *	0.08	0.00	0.36
Choose	0.01	0.02	0.00	**0.04**	**0.02**	**7.28 ****
Match	**0.71**	**0.10**	**76.24** ***	0.38	0.07	0.02
MCQ	**0.05**	**0.16**	**4.45** *	**0.17**	**0.15**	**27.65 *****
Names	**0.03**	**0.16**	**6.96** **	0.08	0.14	0.00
SA	**0.20**	**0.45**	**12.08** ***	0.26	0.59	1.22
Translation	0.00	0.02	0.61	**0.00**	**0.02**	**19.51 *****

*Note: Statistically significant differences are shown in bold. Significance level indicated by *=0.05, **=0.01, ***=0.001.*

Table 8. *Descriptive statistics of item variables in 2017 and 2018 French listening assessments.*

Variable	Foundation			Higher		
	Mean(SE) 2017	Mean(SE) 2018	T-test	Mean(SE) 2017	Mean(SE) 2018	T-test
Discrimination	0.28 (0.02)	0.25 (0.01)	-1.68	0.37 (0.01)	0.36 (0.01)	-0.08
Facility	**0.65 (0.02)**	**0.45 (0.03)**	**-5.39 *****	**0.61 (0.03)**	**0.47 (0.02)**	**4.04 *****
S1 (vocab. diff.)	**1.92 (0.06)**	**2.25 (0.06)**	**3.65 *****	2.98 (0.10)	3.16 (0.07)	1.63
S3 (grammar diff.)	**1.87 (0.08)**	**2.21 (0.07)**	**3.25 ****	2.97 (0.09)	3.15 (0.06)	1.76
S6 (work required)	**1.68 (0.08)**	**2 (0.07)**	**3.04 ****	2.85 (0.10)	2.9 (0.07)	0.45
No. Words	**31.04 (3.26)**	**41.64 (2.43)**	**2.66 ****	59.56 (3.58)	58.42 (2.50)	-0.27
Words per sentence	**8.71 (0.46)**	**9.9 (0.35)**	**2.11 ***	11.92 (0.43)	11.98 (0.33)	0.12
Lexical variety	0.87 (0.02)	0.87 (0.01)	0.03	**0.79 (0.01)**	**0.85 (0.01)**	**4.23 *****
Lexical density	**0.54 (0.01)**	**0.57 (0.01)**	**2.59 ***	**0.55 (0.01)**	**0.57 (0.00)**	**2.49 ***
Lexical familiarity	0.21 (0.01)	0.22 (0.01)	0.67	0.26 (0.01)	0.25 (0.01)	-1.07
Words per second	**1.95 (0.06)**	**1.72 (0.05)**	**-3.09 ****	**1.85 (0.03)**	**1.67 (0.03)**	**-3.78 *****
Track length	**17.59 (2.01)**	**27.85 (1.91)**	**3.64 *****	34.24 (2.34)	37.03 (1.77)	0.96
Pause length	**11.8 (0.65)**	**14.33 (0.65)**	**2.68 ****	**14.48 (0.62)**	**17.7 (0.60)**	**3.6 *****
Time between tracks	32.32 (2.20)	35.20 (2.00)	0.96	39.75 (1.86)	39.11 (1.96)	-0.22

	Proportion 2017	Proportion 2018	Chi-Squared	Proportion 2017	Proportion 2018	Chi-Squared
Instruction language - French	**0.00**	**0.18**	**13.61 *****	**0.00**	**0.23**	**17.66 *****
Picture included	**0.23**	**0.06**	**9.49 ****	**0.16**	**0.00**	**17.22 *****
Topic						
Holidays	0.12	0.20	1.62	0.18	0.11	1.40
Home and environment	0.13	0.15	0.00	0.10	0.18	1.74
Leisure	**0.20**	**0.07**	**5.81 ***	**0.33**	**0.12**	**11.41 *****
Lifestyle	0.32	0.31	0.00	**0.12**	**0.32**	**8.19 ****
Work and Education	0.23	0.27	0.26	0.27	0.28	0.00
Item type						
Blanks	0.00	0.07	3.66	0.00	0.04	1.84
Choose	0.04	0.04	0.00	**0.14**	**0.04**	**3.97 ***
Match	**0.45**	**0.14**	**20.62 *****	**0.18**	**0.04**	**9.22 ****
MCQ	0.24	0.28	0.20	0.26	0.23	0.09
Names	0.11	0.07	0.42	0.11	0.04	1.99
SA	**0.16**	**0.41**	**11.46 *****	**0.32**	**0.60**	**13.46 *****
Gender						
Both	0.11	0.22	3.33	0.21	0.27	0.57
Female	0.48	0.39	1.14	0.41	0.34	0.77
Male	0.41	0.39	0.03	0.38	0.40	0.00

*Note: Statistically significant differences are shown in bold. Significance level indicated by *=0.05, **=0.01, ***=0.001.*

4.4 Multivariate analysis

4.4.1 Facility

The full beta regression models accounting for all the key variables (target language, speech speed or pause length), as well as the additional item and assessment variables, coded proved to fit the data fairly well. Pseudo R-squared scores give an indication of how well the model fits the data, and can range from 0 (models account for no variance in the dependent variable) to 1 (models account for all the variance in the dependent variable). In our models, this statistic increased from 0.05-0.24 in the basic models with just the key variables to 0.44-0.68 in the full models with all of the other variables included, suggesting that the additional variables explain a substantial portion of the variance in facility scores. This suggests that these other features may be of greater importance to item difficulty than target language, speech speed or pause length. Given the much greater fit of the full models, these will be the focus of the discussion below (although details of all models can be found in Appendix 8.3).

4.4.1.1 Effect of key features

Table 9 shows a subset of the results from the full regression models, including just the key features of interest. Speed of speech was a statistically significant factor predicting facility only in the higher tier French listening exams. Modelling suggests that for each additional word per second there is a 76% decrease in the relative probability of a student getting an item completely right (OR = 0.24, CI=0.08-0.7, $p<0.01$). However speed of speech did not have a significant effect on item facility in any of the other assessments. Time between tracks had a small but significant negative effect on facility in the German higher tier assessments (OR=0.98, CI=0.97-0.99, $p<0.05$), suggesting that longer pauses between tracks was related to harder items. Although again this effect was not seen in the other assessments.

The only assessments for which target language had a statistically significant effect on facility after controlling for all of the other item features was in the foundation and higher French reading assessments. In the foundation tier paper, an item in the target language resulted in a 54% reduction (OR=0.46, CI=0.31-0.69, $p<0.001$) in the relative probability of a student answering an item completely correctly (ie it having a facility score of 1). In the higher tier paper an item in the target language resulted in a 34% (OR=0.66, CI=0.46-0.94, $p<0.05$) reduction in the relative probability of a student answering the item completely correctly. To give this some context, if the target language items in these papers had been written in English rather than French, we would predict that the average proportion of students getting the items correct would move from 48% to 64% in the foundation paper and 56% to 65% in the higher paper.

Table 9. *Subset of facility model results, showing odds ratios and significances for key item features.*

		French		German		Spanish	
Paper	Feature	F	H	F	H	F	H
Reading	Instruction language	**0.46*** **	**0.66***	0.70	0.95	1.06	1.04
Listening	Instruction language	1.51	1.15	0.77	1.04	0.65	0.87
	Words per second	0.97	**0.24****	0.71	0.87	0.94	2.21
	Pause length	1.01	0.97	1.03	1.00	0.97	0.99
	Time between tracks	1.00	1.00	1.01	**0.98***	1.00	0.99

*Note: Significance level indicated by *=0.05, **=0.01, ***=0.001.*

4.4.1.2 Additional item and assessment features

In general, the models showed broadly similar patterns in significant predictors across tiers and subjects (summarised in Table 10). Item type was generally a strong predictor of facility with short answer items and translations being the most difficult, and matching type items being the easiest. Question topic was significant in a number of models. Although which topics were easiest or hardest varied substantially between tiers and subjects, the literary extract based items introduced in 2018 to the reading assessments were usually the most difficult.

At least one of the scales scored by the subject experts was significant in all models. The complexity of the vocabulary and the 'work required' by students in the majority of models had a strong negative impact on item facility (see Appendix C for details of model results). Lexical variety was also linked to more difficult items in a number of models such that items with more unique words tended to be harder, although this effect was not consistent across models.

Table 11 indicates what are likely to be the main drivers of lower facility scores overall in 2018. It combines the results from the facility regression models and the identification of which features have changed between 2017 and 2018. Cells have been shaded to indicate if the shift in that feature (either positive or negative) is likely to have contributed to making the overall assessment easier (shaded green) or more difficult (shaded red).

Table 10. *Summary of facility model results.*

	French				German				Spanish			
	Reading		Listening		Reading		Listening		Reading		Listening	
	F	H	F	H	F	H	F	H	F	H	F	H
Instruction language	−	−										
Speed of speech				−								
Pause length												
Time between tracks								−				
S1 (Vocab. difficulty)	−	−		−	−	−		−			−	
S3 (Grammar difficulty)			−									
S6 (Work Required)	−					−	−		−	−	−	−
Word count		−			+				−	−		
Words per sentence											−	−
Lexical variety	−	−						−	−	−		
Lexical density												
Lexical unfamiliarity												
Pictures	+		+						+	−		
Topic	Y		Y	Y		Y	Y			Y		
Item type	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Track length				−								
Gender of speaker(s)			Y									
Pseudo R-Squared	**0.61**	**0.44**	**0.66**	**0.47**	**0.60**	**0.54**	**0.68**	**0.51**	**0.65**	**0.59**	**0.62**	**0.64**

Note: +/- indicates if this feature had a significant positive or negative impact on facility. Y indicates a categorical variable where at least one category had a significant effect on facility.

Table 11. *Key drivers of the change in facility between 2017 and 2018.*

	French				German				Spanish			
	Reading		Listening		Reading		Listening		Reading		Listening	
	F	H	F	H	F	H	F	H	F	H	F	H
Instruction language												
Speed of speech												
Pause length												
Time between tracks												
S1 (Vocab. difficulty)												
S3 (Grammar difficulty)												
S6 (Work Required)												
Word count												
Words per sentence												
Lexical variety												
Lexical density												
Lexical unfamiliarity												
Pictures												
Topic												
Item type												
Track length												
Gender of speaker(s)												
Pseudo R-Squared	**0.61**	**0.44**	**0.66**	**0.47**	**0.60**	**0.54**	**0.68**	**0.51**	**0.65**	**0.59**	**0.62**	**0.64**

Note: Cells are shaded based on whether this feature had a significant impact on item facility and showed an increase or decrease between 2017 and 2018. Red shading indicates that the changes are estimated to increase difficulty, green indicates the changes are estimated to reduce difficulty of the reformed assessments.

Overall, Table 11 indicates which changes are associated with a change in difficulty between 2017 and 2018. The shift in item type use has a strong consistent impact on difficulty. The translation items introduced in the 2018 reading assessments are generally the most difficult and there has been an increase in the use of short answer items which are the most difficult item type in most listening assessments (see Appendix 8.3 for details). This change in the frequency of item types is therefore likely to be a major contributor to the increased assessment difficulty in 2018. Another feature with the most consistent impact on difficulty is the increase in more demanding vocabulary (S1 – vocabulary difficulty) and an increase in the difficulty for students to identify the information required to answer each question (S6 – work required). The change in S6 (work required) is likely due to fewer items allowing the identification of single key words and a greater requirement for students to comprehend complete passages, essentially requiring more ‘work’ from students.

This is also likely linked to the change in item types, with the reduction in the use of straightforward matching type questions.

Full model results also suggest that even after including all of the variables in the model, there is still a significant effect of exam board and/or year in some of the models. This suggests that there are differences between the exam boards which are affecting facility scores but that are not controlled for. This could be due to features of the assessments which differ between exam boards but are not sufficiently accounted for in the models. This means that the effects we estimated have to be considered as average effects. If particular features are significantly more prevalent in one exam board's assessments than another's this could affect the estimated effect of these features on item difficulty. The presence of unaccounted factors for differences between years also indicate that some of the differences in facility over time are not controlled for. This could be due to features of the assessments that have changed with the reforms, differences in students' ability or familiarity with the assessments (sawtooth effect). Also in this case, therefore, the effects estimated by our models have to be interpreted as average effect over time.

4.4.2 Discrimination

Discrimination is the other key aspect of item functioning, which gives an indication of how well each item differentiates between students of differing ability. Similar analysis was ran to the facility modelling but with discrimination score as the dependent variable and using a linear regression rather than a beta regression models, as discrimination scores were normally distributed. Analysis suggested that none of the features of interest had a very strong or consistent impact on discrimination between assessments. There is some indication that questions in the target language improve discrimination, however this was only statistically significant in two Spanish assessments (Table 12). Longer pauses may also have a slight negative effect on discrimination in German higher listening assessments.

Among all of the item features included, the most salient feature affecting discrimination was item type (See Appendix D for full model output). In the listening assessments, items requiring filling in blanks were usually the least discriminating items. In the reading assessments, multiple choice items or matching-type items were consistently the least discriminating items. By far the best discriminating items were the translation items in the reading assessments, although this may be due to them being the only items awarding over 2 marks, therefore allowing better differentiation (see Appendix D). Overall these changes in item types may explain the slight increase in average discrimination in 2018.

Table 12. *Key subset of full discrimination model results for key item features, showing beta estimates.*

		French		German		Spanish	
Paper	Feature	F	H	F	H	F	H
Reading	Instruction language	0.02	0.04	-0.04	0.01	**0.05***	0.00
Listening	Instruction language	-0.01	0.03	-0.01	0.03	0.06	**0.07***
	Words per second	0.04	0.14	0.03	0.09	0.06	-0.04
	Pause length	0.01	0.01	0.00	**-0.01*****	0.00	0.00
	Time between tracks	0.00	0.00	0.00	**0.00****	0.00	0.00

*Note: Significance level indicated by *=0.05, **=0.01, ***=0.001.*

5 Results – Evaluation of overall assessment outcomes

The second aim of this study was to identify how the overall assessments functioned, particularly by observing how the relationship between the different assessment components has changed with the reforms and to consider whether there is any evidence that students have been disadvantaged. Initially we present the subject level outcomes, then the component level outcomes. We then look at the relationship between the different components in 2017 and 2018 and the results of a logistic model to predict student outcomes on the different components to identify if there are substantial changes in the difficulty of components between years.

5.1 Subject level outcomes

Table 13 shows that subject level outcomes have remained relatively stable between 2017 and 2018 in all cases. This is due to the standard setting methodology used for the first awards of reformed GCSEs, which is designed to ensure that students are not disadvantaged by being the first to sit new qualifications. To compensate for the increase in difficulty in 2018 and therefore reduction in marks achieved (see Appendix E for mark distributions), grade boundaries are lower across all three exam boards offering MFL assessments (see Appendix F for details of grade boundary changes). This approach was used in the transition to reformed GCSEs so that students in 2018 were, on average, as likely as students in 2017 showing similar prior attainment to achieve a grade C/4 (or A/7) and above.

Table 13. *Proportion of* student*s attaining C/4 and above and A/7 at subject level in 2017 and 2018.*

Subject	Total Entry		Percentage C/4 and above		Percentage A/7 and above	
	2017	2018	2017	2018	2017	2018
French	106416	115505	22.2%	22.5%	69.4%	69.1%
German	36876	40967	22.2%	22.1%	74.6%	74.5%
Spanish	74005	86075	25.9%	25.9%	70.0%	69.6%

5.2 Component level outcomes

Due to the data available, component level analysis was restricted to listening, reading and writing components.

Table 14 shows the percentage of students obtaining a C/4 or above in the foundation and higher tier papers. Table 15 shows the percentage of students obtaining an A/7 or above in the higher tier papers[6]. Tables are shaded to indicate where proportions have increased (green) or decreased (red) in 2018.

The percentage of foundation tier students achieving a C/4 or above on the writing assessment is lower in 2018, however attainment on the listening and reading assessments is higher. On the higher tier, the proportion gaining a C/4 or above in writing has remained fairly stable, but is higher for the listening and reading components. Also on the higher tier, the proportion of students attaining A/7 and above in listening and reading is higher, whereas in writing it is lower. In general this suggests a more even distribution of grades across these three components in 2018 than 2017, which should provide a better spread of marks at qualification level and ensures that each skill contributes equally to the overall qualification grade.

Table 14. *Percentage of students attaining C/4 and above, by tier and component*

Tier	Components	French		German		Spanish	
		2017	2018	2017	2018	2017	2018
Foundation	Listening	9%	44%	13%	32%	13%	40%
	Reading	12%	49%	21%	55%	21%	38%
	Writing	68%	41%	60%	50%	63%	41%
Higher	Listening	66%	91%	66%	95%	71%	87%
	Reading	75%	95%	71%	95%	68%	88%
	Writing	95%	90%	91%	91%	93%	90%

Note: Red shading indicates that the percentage is lower in 2018 than 2017, green indicates the percentage is higher in 2019 than 2018.

For 2017 tier was defined by which tier students took the listening and reading components in, and only included students who took both the listening and reading components in the same tier.

[6] In 2018, as all the qualifications are linear, component grades are notional and give an indication of candidate performance but play no part in the determination of qualification grades.

Table 15. *Percentage of students attaining A/7 and above, by component (higher tier only)*

Components	French		German		Spanish	
	2017	2018	2017	2018	2017	2018
Listening	37%	42%	27%	41%	39%	46%
Reading	45%	48%	26%	48%	48%	48%
Writing	53%	46%	40%	37%	52%	47%

Note: Red shading indicates that the percentage is lower in 2018 than 2017, green indicates the percentage is higher in 2019 than 2018.

For 2017 tier was defined by which tier students took the listening and reading components in, and only included students who took both the listening and reading components in the same tier.

5.2.1 Relationship between components

Table 16 shows the correlations between students' marks in each component. These give an indication of the relationship between the components and to what extent they are measuring the same underlying trait or ability. Generally we would expect different assessments within the same subject to be reasonably well correlated. One of the reasons that controlled assessment was removed for writing in the reformed specifications was its poor ability to differentiate between students, as many students received high marks. The knock on effect of this was that grade boundaries were relatively high for the other components to compensate for students' generally high marks in the writing unit. In the reformed specifications the move to assessing writing through an exam aimed to bring it more in line with the reading and listening components.

Table 16. *Correlation coefficients of student standardised marks between different components.*

Tier	Components		French		German		Spanish	
			2017	2018	2017	2018	2017	2018
Foundation	Listening	Reading	0.7	0.51	0.53	0.3	0.72	0.72
	Reading	Writing	0.27	0.78	0.31	0.56	0.34	0.74
	Writing	Listening	0.26	0.49	0.26	0.64	0.36	0.66
Higher	Listening	Reading	0.78	0.82	0.73	0.74	0.81	0.83
	Reading	Writing	0.3	0.68	0.43	0.75	0.4	0.75
	Writing	Listening	0.2	0.55	0.36	0.73	0.27	0.65

Note: All correlations significant at the p<0.001 level.

Table 16 indicates that the correlation between the writing assessment and the other two assessments is much higher in 2018 compared to 2017 in all three languages and across both tiers. For example, the correlation between the reading and writing assessments for foundation tier French was 0.78 in 2018, compared to 0.27 in 2017. This may be in part related to the assessment now taking a similar exam format, and therefore students are demonstrating a similar set of exam skills. However, the exam has also allowed better differentiation of students in the marks achieved in 2018, which will contribute to a higher correlation with the other assessments (see Appendix 8.5). The lower correlation between the French and German reading and listening assessments is less easy to explain, but may be due to the foundation tier listening assessments in these two languages having shifted to become more difficult, whereas the corresponding reading assessment has seen less of a shift in 2018. In future years we would therefore expect this correlation to increase as students and teachers become more familiar with the reformed listening assessments.

5.2.2 Relationship between component and subject level grades

To explore whether these changes in outcomes provide a better indication of student ability, the relationship between each component outcome and qualification outcome is shown below. This essentially gives us an indication of whether components effectively differentiate between students of different ability. Table 17 shows the percentage of students who attained A/7 and above and C/4 and above in each component who went on to attain an A/7 and above at qualification level. Similarly, Table 18 shows those who obtained C/4 and above at qualification level. If all of the assessments are equally contributing to qualification outcomes we would expect the percentages across the components to be similar. Generally in each case we would expect students who achieved A/7 and above in a component to have a greater percentage chance of obtaining A/7 or C/4 (and above) overall than a student who attained a C/4 in each component. We would subsequently expect that students obtaining a C/4 and above in an individual component would have a moderate chance of obtaining at least a C/4 overall and a significantly lower chance of obtaining at least an A/7 overall, if the assessment differentiates well.

Table 17 and Table 18 indicate that both the reading and listening assessments are a better predictor of overall outcomes in 2018, which is likely to be due to their increased relative contribution to the qualification grade, given the difference in component weightings in 2018. Gaining at least an A/7 in the writing assessment is a better predictor in 2018 of attaining an A/7 overall. However, students obtaining at least a C/4 in writing are less likely to get at least an A/7 or a C/4 in 2018 than 2017. This may be due to the reduced contribution of the writing assessments to the overall qualification grade, but may also suggest that the writing assessment now gives a

better reflection of overall student ability and is more in line with the other components, as was seen in the previous section.

Table 17. *Percentage of students who obtained an A/7 or above and C/4 or above in each component who attained an A/7 or above at qualification level.*

		French		German		Spanish	
		2017	2018	2017	2018	2017	2018
Reading	A/7+	83.2%	90.6%	65.5%	94.7%	87.7%	88.6%
	C/4+	33.4%	37.5%	22.5%	39.2%	40.0%	38.1%
Listening	A/7+	69.9%	84.2%	62.2%	100.0%	73.6%	82.3%
	C/4+	26.1%	32.8%	23.5%	88.8%	31.9%	37.1%
Writing	A/7+	81.1%	85.6%	78.6%	81.5%	78.7%	83.1%
	C/4+	51.0%	35.3%	42.5%	29.6%	53.5%	37.8%

Table 18. *Percentage of students who obtained an A/7 or above and C/4 or above in each component who attained an C/4 or above at qualification level.*

		French		German		Spanish	
		2017	2018	2017	2018	2017	2018
Reading	A/7+	99.0%	100.0%	98.9%	100.0%	98.7%	99.9%
	C/4+	65.0%	94.7%	69.8%	96.2%	66.9%	89.4%
Listening	A/7+	95.9%	99.9%	93.1%	86.9%	96.3%	99.7%
	C/4+	54.9%	89.2%	61.9%	33.5%	65.9%	87.3%
Writing	A/7+	99.7%	99.7%	99.9%	100.0%	99.6%	99.9%
	C/4+	97.0%	90.9%	94.7%	93.5%	96.0%	92.0%

5.3 Relative component difficulty

An analysis of relative component difficulty in 2017 and 2018 was carried out using a series of logistic regression models. In these models the likelihood of achieving a C/4 or above (Table 19), or A/7 or above (Table 20) was predicted using students' prior attainment, the identity of the component, the year and an interaction between year and component as independent variables. The figures in the tables show odds ratios, essentially what we would expect to multiply the probability of the dependent variable by for each unit change of the independent variable. For prior attainment this means that a value over 1 indicates the expected relative percentage increase in the probability of attaining the grade in question for each point higher a student achieved in their mean KS2 score. For example, a value of 1.08 would indicate an 8% increase in the relative probability of achieving the grade in question for each additional increase in the students' KS2 prior attainment score. For the different

assessments the odds ratios use the writing component as a reference, so each odds ratio explains the difference in the relative probability of attaining the grade in question in listening or reading when compared to the probability for writing. Therefore a value over 1 indicates that reading and listening were easier, whereas a value under 1 indicates that writing was easier. The figures presented alongside show the same information in a different way.

For each KS2 score on the x-axis the probability of attaining at least a C/4 or at least an A/7 in each assessment can be estimated by looking at the relative position of each curve on the y-axis. If the curves are close together then the probability between each assessment is similar, if the curves are far apart then the line higher up indicates an easier assessment and the line lower down indicates a more difficult assessment.

The model indicated that in 2017 it was significantly harder to obtain a C/4 or above in the listening and reading component than in the writing component. This is shown by the odds ratios for listening and reading compared to writing being significantly below 1 (and to a lesser extent A/7; see figures 7 and 8). In 2018 the likelihood of attaining a C/4 in writing was reduced, however the likelihood of attaining at least a C/4 in listening and reading increased, with the difference in difficulty of the assessments being much reduced across the grade range. This is indicated by the odds ratios being closer to 1 in 2018 and the lines in the figures being closer together. This pattern was similar across the three languages (see Appendix G for Spanish and German).

In 2017, an average French student was 93% less likely to attain at least a C/4 in listening than writing and 89% less likely in reading than writing. In 2018, an average student was only 7% *more* likely to obtain at least a C/4 in listening than writing and 47% *more* likely in reading than writing. These patterns are broadly similar across German and Spanish assessments. In 2017 writing was consistently the easiest assessment. In 2018 across all languages, writing has been brought closer in line with the other assessments in terms of the probability for a student with a similar prior attainment to achieve at least a C/4 or at least an A/7.

Table 19. *Odds ratios of model results for the probability of attaining C/4 or above.*

	French		German		Spanish	
	2017	2018	2017	2018	2017	2018
Prior attainment	**1.08*** **	**1.08*** **	**1.07*** **	**1.08*** **	**1.07*** **	**1.07*** **
Skill [Writing]						
Listening	**0.07*** **	**1.07*** **	**0.15*** **	**0.72*** **	**0.15*** **	**0.89*** **
Reading	**0.11*** **	**1.47*** **	**0.20*** **	**1.48*** **	**0.15*** **	**0.88*** **

*Note: Significance level indicated by *=0.05, **=0.01, ***=0.001.*

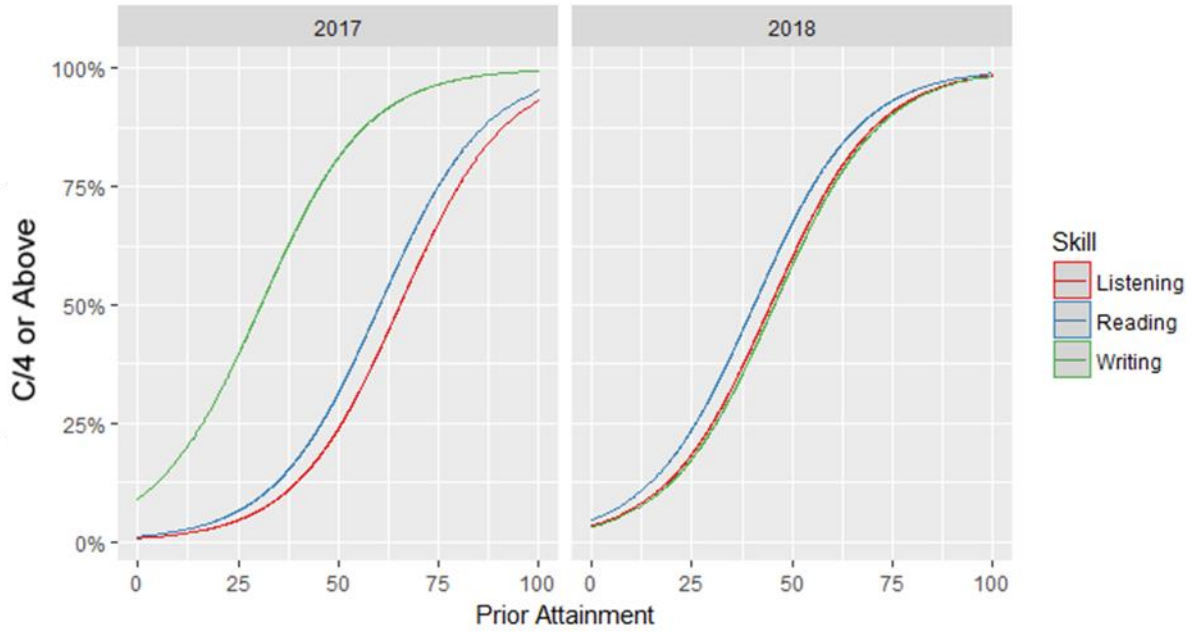

Figure 7. *Probability of attaining a C or above in different components by prior attainment in French.*

Table 20. *Odds ratios of model results for the probability of attaining A/7 or above.*

	French		German		Spanish	
	2017	2018	2017	2018	2017	2018
Prior attainment	**1.07*** **	**1.09*** **	**1.07*** **	**1.09*** **	**1.06*** **	**1.08*** **
Skill [Writing]						
Listening	**0.29*** **	**0.83*** **	**0.36*** **	**1.20*** **	**0.34*** **	**0.92*** **
Reading	**0.43*** **	**1.06*** **	**0.35*** **	**1.63*** **	**0.51*** **	0.97

*Note: Significance level indicated by *=0.05, **=0.01, ***=0.001.*

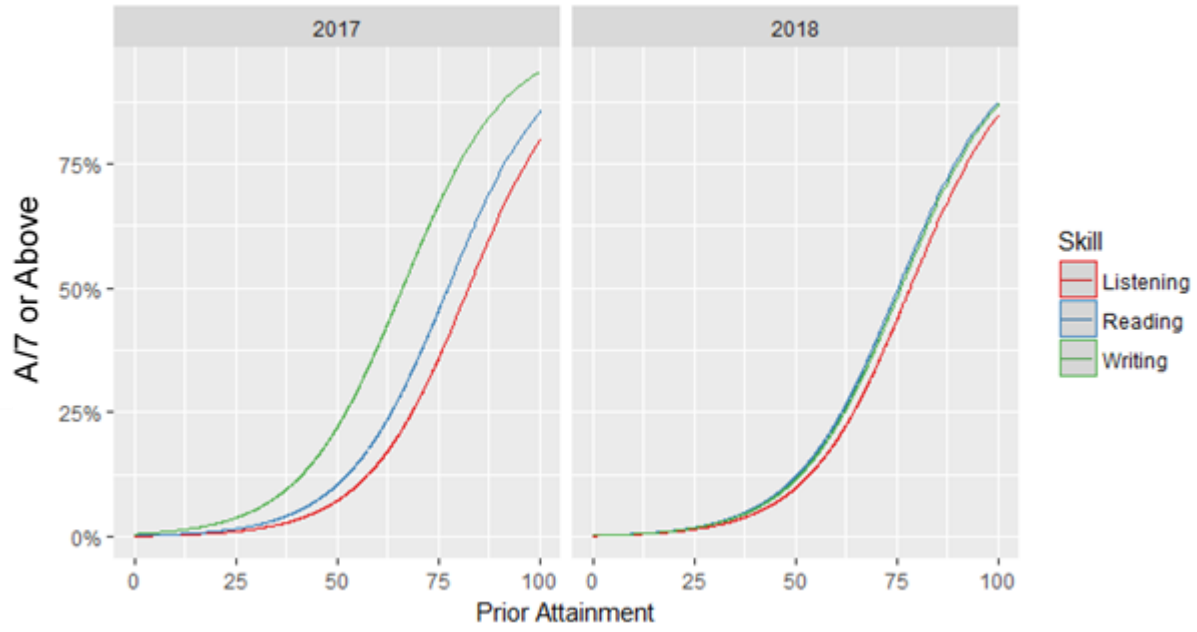

Figure 8. *Probability of attaining an A/7 or above in different components by prior attainment in French.*

The analysis was repeated but split by tier to investigate if the pattern between assessments was similar for students across the grade range. When split by tier (using the tier for the listening and reading assessments to assign tier for writing in 2017), the difference between writing and the other assessments are more pronounced in the foundation tier (see figures 9 and 10). For French assessments, in the foundation tier in 2017 students were 96% less likely to obtain at least a C/4 in listening than writing components and 94% less likely to obtain a C/4 or above in reading than writing, whereas in 2018 students were 13% more likely to obtain at least a C/4 in listening than writing and 44% more likely to obtain at least a C/4 in reading (Table 21). This is comparable to the higher tier where in 2017 the relative probability of attaining at least a C/4 on listening was only 57% lower for listening than writing and only 35% lower for reading than writing. Whereas in 2018 there was only a 19% lower relative probability of obtaining a C in listening than writing and were 8% more likely to obtain a C in reading than writing (Table 21 and 22).

Again these patterns are broadly similar across the languages with the gap between writing and the other assessments being greater in the foundation tier than the higher tier and in all cases lower in 2018 compared to 2017. This suggests that the change in the writing assessment has consistently brought it more in line with reading and writing in terms of the probability of obtaining at least a C/4 in 2018.

Table 21. *Odds ratios of model results for the probability of attaining C or above, foundation tier.*

	French		German		Spanish	
	2017	2018	2017	2018	2017	2018
Prior attainment	**1.04*** **	**1.05*** **	**1.03*** **	**1.04*** **	**1.03*** **	**1.04*** **
Skill [Writing]						
Listening	**0.04*** **	**1.13*** **	**0.09*** **	**0.44*** **	**0.08*** **	**0.96***
Reading	**0.06*** **	**1.44*** **	**0.17*** **	**1.04*** **	**0.15*** **	**0.89*** **

*Note: Significance level indicated by *=0.05, **=0.01, ***=0.001.*

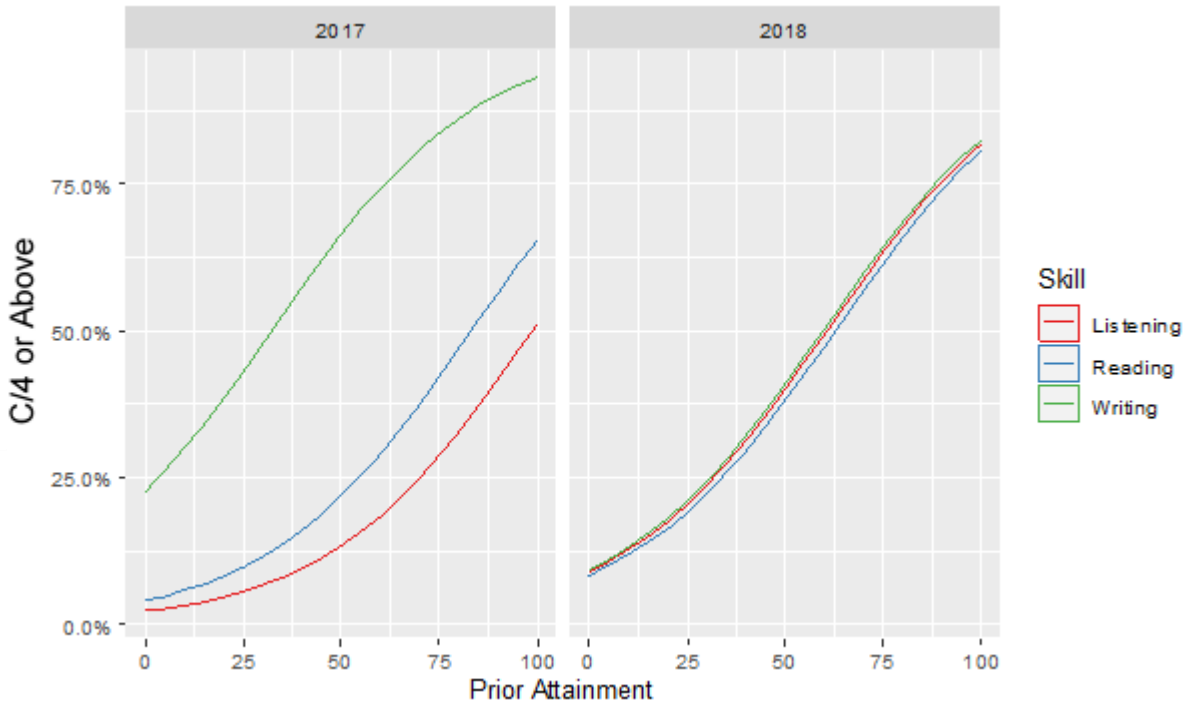

Figure 9. *Probability of attaining a C or above in different components by prior attainment in French, foundation tier only.*

Table 22. *Odds ratios of model results for the probability of attaining C or above, higher tier.*

	French		German		Spanish	
	2017	2018	2017	2018	2017	2018
Prior attainment	**1.05*** **	**1.06*** **	**1.06*** **	**1.06*** **	**1.05*** **	**1.04*** **
Skill [Writing]						
Listening	**0.43*** **	**0.81*** **	**0.49*** **	**1.27*** **	**0.5*** **	**0.90*** **
Reading	**0.65*** **	**1.08*** **	**0.46*** **	**1.71*** **	**0.78*** **	0.97

*Note: Significance level indicated by *=0.05, **=0.01, ***=0.001.*

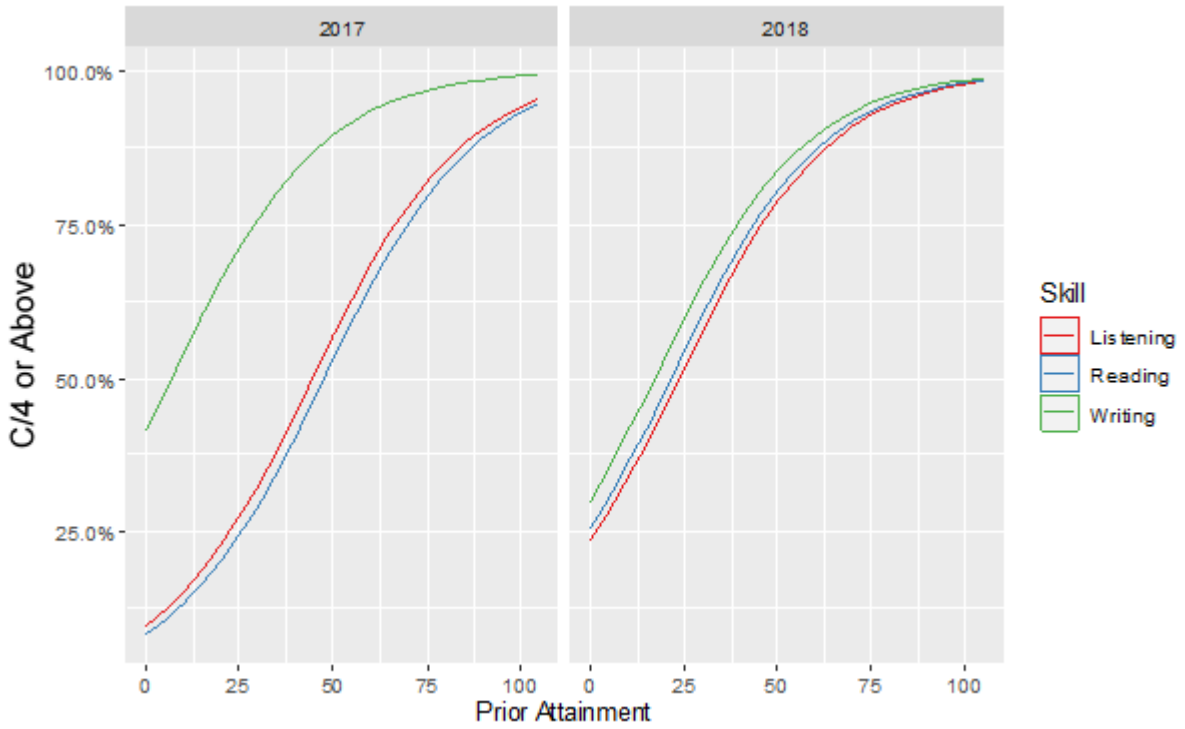

Figure 10. *Probability of attaining a C or above in different components by prior attainment in French, higher tier only.*

6 Discussion

This aim of this report was to evaluate the changes to GCSE MFL assessments, by assessing whether reformed assessments were fair to students and identifying whether there was any evidence that the recent reform had disadvantaged students taking the assessments in 2018 rather than in 2017. In order to answer this question, two strands of work were undertaken. First, we thoroughly analysed the item features that affect difficulty, focussing on those that were changed in the reformed specifications. Second, we considered how the relationships between students' performance on the different elements of the assessment have changed between 2017 and 2018.

6.1 Summary of findings

The findings suggest that the reformed assessments in 2018 are functioning better than the pre-reform assessments. Analysis shows that the mean facility scores have generally decreased in 2018 indicating an overall increase in difficulty. This effect was relatively consistent across languages, skills, tiers and exam boards. Although difficulty may have increased in 2018, analysis suggest that in most cases this is likely to have had a positive impact on the classification accuracy of students. Due to the approach taken to carry forward standards in the reformed GCSEs, students were not disadvantaged because of the increase in difficulty, and the proportion of students at each grade in 2018 was similar to 2017. Reformed assessments produced an increase in the spread of marks and therefore allowed the spreading out of grade boundaries, giving greater confidence that students are receiving the grade their work deserves. Discrimination analysis also suggests that, at least in French and German, there has been a slight average improvement in item level discrimination, which means that items in these assessments are, on average, slightly better at creating a consistent rank order of students by ability.

The increase in difficulty between 2017 and 2018 assessments is likely not due to those features which were initially of concern to stakeholders. Speed of speech and pause length in the listening assessments had little effect on item facility and did not change substantially between years. The introduction of questions in the target language only had a significant impact on item facility in the French reading assessment, which may have contributed to the increased difficulty in 2018, but not to a degree where items would likely become inaccessible. Although this is an aspect to be considered by exam boards and item writers in the future, it suggests that these items are not overly difficult for the cohort taking the assessments.

Our models indicated that the features which had the biggest impact on change in facility in 2018 were an increase in the demand of the vocabulary used in the reading and listening texts, and items requiring more 'work' from students to answer the

question (for instance, not being able to rely on spotting key words or phrases). As an aside, the predictive power of these features indicates that utilising subject experts proved to be an effective way of holistically considering item features without requiring highly complex models of linguistic features.

The introduction of the extract-based items, translation items and more short answer items is likely to have also increased difficulty. These changes are in line with the intentional increase in the demand stipulated by DfE as part of the reforms to GCSEs. Arguably items in 2018 may be a more valid reflection of student's ability in the target language, given that the key causes of increased difficulty were due to vocabulary demand and requiring more work from students, which are likely to be closely related to the language ability construct. A detailed analysis of the validity of the content of these assessments has been carried out in a separate study (Curcin & Black, 2019).

In a couple of cases, the facility scores for the 2018 assessments were quite low (potentially suggesting these assessments were too difficult). However, it is possible that this was due to teachers and students lack of familiarity with the new assessments.

The changes to the writing assessment have improved the balance between the assessments, in terms of the weighting compared to the other assessments and the distribution of marks. In 2017 students' marks on the writing assessment were a poor predictor of attainment on the reading and listening assessments. It was also much easier to obtain at least a C/4 in the writing assessment than in the other components. In 2018, the components were better balanced. The decrease in attainment in the writing assessment has been balanced by an increase in attainment in both the reading and listening components. This is particularly noticeable for foundation tier students where, in the previous assessments, the score in the writing assessment was substantially higher than the other assessments. This change in balance has allowed better differentiation in the listening and reading assessments as notional component grade boundaries in these assessments are lower in 2018 than in 2017. However due to the methodology used to maintain standards, qualification level outcomes were stable. This means that qualification outcomes better represent students' ability across the different skills.

6.2 Limitations and further research

Our statistical models of facility scores did not account for all variation in difficulty between items and assessments. Although they did explain a good proportion of the variance, there is still a large proportion unexplained. The models used here included assessment features which could be relatively easily scored or rated. However, it is likely that there are other more subtle features affecting the difficulty of items and assessments overall. In particular, it is likely that different features may

interact in potentially complex ways, which were not accounted for in these models. Our regression models were kept intentionally relatively simple and although this might have precluded our ability to capture some subtleties, this has the advantage of allowing us to explain the complexity of the problem (ie how item features affect difficulty) in a relatively straightforward way.

Given that the statistical models did not predict facility with a 100% accuracy, we attempted to identify some of these factors which affected item difficulty but which had not been accounted for. Subject experts were presented with a series of items for which the facility values were poorly predicted by the statistical models. The experts were then tasked to comment on any features of each item which may have caused them to be more or less difficult than expected, and which were not already included in the models. We collated these insights and summarised them across subject experts and languages. In general, where items which were *easier* than predicted, subject experts thought this was due to:

- Answers which allow lifting words or sentences straight from provided text;
- Generous mark schemes;
- Cognate words which look or sound similar in English being key to the answer;
- Guessable answers for multiple choice type questions.

Where items were *more difficult* than expected, the subject experts thought this was due to:

- Overly restrictive mark schemes;
- Difficult distractors for multiple choice type items;
- Misleading sections of text, which led students to give the wrong answer;
- Poorly written questions, in some cases with confusing wording in English;
- Difficult synonyms used in the text, requiring interpretation or inference as a direct translation was not available;
- Distracting or confusing voice acting for the listening assessments.

Unfortunately these item features were difficult to encode into the statistical models, but do help explain some of the unaccounted for variability in item difficulty. Further research may be needed to try and account for these aspects more systematically. In the meanwhile, however, these findings will be shared and discussed with exam boards so that they will be able to take them into account in the development of future assessments.

The use of facility as a dependent variable in the key models has weaknesses as it is inherently related to the ability of the cohort. Although we tried to account for this by including measures of concurrent ability, exam board and year within the models

there are potential differences in the cohort not accounted for by these measures. This could potentially have distorted our results if certain features linked to facility are, for example, also linked to a specific exam board and this exam board has a slightly different cohort of students. However, as we were generally looking for broad patterns across assessments and languages it is unlikely that these distortions would fundamentally change our conclusions.

It should be noted that the approach used in this report is meant to produce evidence on the relationship between certain item features and difficulty. This approach does not allow us to address the concerns raised by some stakeholders as to whether, as an example, the use of vocabulary in the assessment is appropriate. Further research may be needed to look at the validity of certain item features such as vocabulary use.

A final limitation of our analysis, which has been alluded to elsewhere, is that it only covered the first year of reforms. Some features of these assessments were new to both students and teachers in 2018. Previous research has indicated that it can take up to three years for the effect of exam familiarisation to cease having an impact on assessment outcomes (Cuff *et al.,* 2019). Particularly for the new item types (translation, literary extracts) and the new assessment in writing, the lower facility scores may have been, at least in part, due to lack of familiarity. Therefore, if this analysis were to be rerun in 2019 or beyond, it is possible that different features would be flagged as having a greater influence over item difficulty.

6.3 Conclusion

Within the limitations discussed above, the analysis presented here was successful in identifying key factors related to the difficulty of the MFL assessments and the likely causes of an increase in difficulty in the reformed assessments were identified. Concerns which were raised prior to the assessments being sat which were a key focus of this evaluation (ie speed of speech, pause length, target language questions) were found to not have a negative impact on the students taking these assessments. Due to the approach taken to carrying forward standards from the legacy to the reformed specifications, any change in assessment difficulty did not result in lower qualification outcomes. Overall, from a technical perspective, the reformed assessments are functioning better than the pre-reform assessment. When combined with an increased balance among components, the net result is greater differentiation between students, with GCSE grades better representing students' ability across the range of skills. The findings of this report point towards the conclusion that, although there may be still some room for improvement in some aspects of the assessments, we can be confident that the reformed assessments did not disadvantage students sitting GCSE MFL in 2018 and in fact provided a fairer representation of their knowledge and skills.

7 References

Ahmed, A., & Pollitt, A. (1999). *Curriculum demands and question difficulty.* Paper presented at the International Association for Educational Assessment, Bled.

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension.* Maryland University College Park.

Board, K. and Tinsley, T. (2016). *Language Trends 2015/16. The state of language learning in primary and secondary schools in England.* Reading: Education Development Trust. Retrieved from: https://www.britishcouncil.org/sites/default/files/language_trends_survey_2016_0.pdf

Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing, 23(3)*, 269-289.

Churchward, D. (2019). *Recent trends in modern foreign language exam entries in anglophone countries*. (Report No. Ofqual/19/6557). Coventry, UK: Ofqual.

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.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research, 48(2),* 139-154.

Crocker, L., & Algina, J. 2008. *Introduction to classical and modern test theory*. Ohio; United Kingdom: Wadsworth.

Cuff, B. M., Meadows, M., & Black, B. (2019). An investigation into the Sawtooth Effect in secondary school assessments in England. *Assessment in Education: Principles, Policy & Practice, 26(3),* 321-339.

Curcin, M., & Black, B. (2019). *Investigating standards in GCSE French, German and Spanish through the lens of the CEFR*. (Report No. Ofqual/19/6559). Coventry, UK: Ofqual.

DfE (2015). *Modern languages GCSE subject content*. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/485567/GCSE_subject_content_modern_foreign_langs.pdf

El Masri, Yasmine H., Ferrara, Steve, Foltz, Peter W., & Baird, Jo-Anne. (2017). Predicting Item Difficulty of Science National Curriculum Tests: The Case of Key Stage 2 Assessments. *Curriculum Journal, 28(1),* 59-82.

Fisher-Hoch, H., Hughes, S., & Bramley, T. (1997). *What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions.* Paper

presented at the British Educational Research Association Annual Conference, University of York.

Guardian, 2015. https://www.theguardian.com/teacher-network/2015/aug/20/why-drop-students-langauge-gcses-teachers-views

Guardian, 2019. https://www.theguardian.com/education/2019/may/11/modern-language-teaching-under-threat-from-tough-exams

TES, 2019. https://www.tes.com/news/tougher-gcses-put-students-learning-languages

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and Validating Test Items*. New York: Routledge.

Jadhav, C. (2018, 6 February). *New GCSEs in French, German and Spanish.* Ofqual Blog. Retrieved from: https://ofqual.blog.gov.uk/2018/02/06/new-gcses-in-french-german-and-spanish/

JCQ (2016). *GCE Trends 2016.* Retrieved from: http://www.jcq.org.uk/examinationresults/a-levels/2016/gce-trends-2016

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics, 16(3),* 307-322.

Lumley, T., Routitsky, A., Mendelovits, J & Ramalingam, D. (2012). *A framework for predicting item difficulty in reading tests*. Retrieved from: http://research.acer.edu.au/pisa/5

Ofqual (2013). *Review of Controlled Assessments in GCSEs*. (Report No. Ofqual/13/5291). Coventry, UK: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/377903/2013-06-11-review-of-controlled-assessment-in-GCSEs.pdf

Ofqual (2016). *An investigation into the 'Sawtooth Effect' in GCSE and AS / A level assessments*. (Report No. Ofqual/16/6098). Coventry, UK: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/549686/an-investigation-into-the-sawtooth-effect-in-gcse-as-and-a-level-assessments.pdf

Ofqual (2017). *GCSE Subject Level Conditions and Requirements for Modern Foreign Languages.* (Report No. Ofqual/17/6161). Coventry, UK: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/592158/GCSE_Subject_Level_Conditions_for_Modern_Foreign_Languages_Feb_2017.pdf

Ofqual (2018). *Inter-subject comparability in A level sciences and modern foreign languages*. (Report No. Ofqual/18/6450). Coventry, UK: Ofqual. Retrieved from:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/757841/ISC_Decision_Document_20.11.18.pdf

Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H., & Bramley, T. (1998). *The effects of structure on the demands in GCSE and A level questions*. Report to Qualifications and Curriculum Authority. University of Cambridge Local Examinations Syndicate.

Pollitt, A., Entwistle, N., Hutchinson, C., & De Luca, C. (1985*). What makes exam questions difficult?* Edinburgh: Scottish Academic Press.

Pollitt, A., Ahmed, A., & Crisp, V. (2007). *The demands on examination syllabuses and question papers*. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), Techniques for monitoring the comparability of examination standards (pp. 166–206). London: Qualifications and Curriculum Authority

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing, 1(3 & 4),* 185–216.

Stratton, T. (2019, 14 February). *Evaluating new GCSEs in French, German and Spanish.* Ofqual Blog. Retrieved from: https://ofqual.blog.gov.uk/2019/02/14/evaluating-new-gcses-in-french-german-and-spanish/

Taylor, R. and Opposs, D. (2018). '*Standard setting in England: A leve*ls'. In Baird, J., Isaacs, T., Opposs, D. and Gray, L. (Eds) Examination standards: how measures & meanings differ around the world. London: UCL IOE Press

Taylor, R. and Zanini, N. (2017). *Native speakers in A level modern foreign languages.* (Report No. Ofqual/17/6203). Coventry, UK: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/610147/Native_speakers_in_A_level_modern_foreign_languages.pdf

Tinsley, T., & Doležal, N. (2018). *Language Trends 2015/16. The state of language learning in primary and secondary schools in England.* Reading: Education Development Trust. Retrieved from: https://www.britishcouncil.org/sites/default/files/language_trends_2018_report.pdf

Ure, J. (1971). *Lexical density and register differentiation*. In G. Perren, J.L.M. Trim (Eds.). Applications of Linguistics. Selected Papers of the Second International Congress of Applied Linguistics, Cambridge 1969.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2011). Acquiring item difficulty estimates: a collaborative effort of data and judgment. In *Proceedings of the 4th international conference on educational data mining* (pp. 121-127).; Eindhoven: Eindhoven University of Technology.

Appendix A – Overall descriptive statistics

Table A.1. Descriptive statistics for reading assessments.

Reading	French				German				Spanish			
	Foundation		Higher		Foundation		Higher		Foundation		Higher	
	2017	2018	2017	2018	2017	2018	2017	2018	2017	2018	2017	2018
No. Students	43,595	52,114	66,182	60,396	10,638	14,138	26,585	23,531	26,843	36,992	49,182	46,388
Mean KS2	48.75	49.00	62.82	64.12	49.86	50.56	63.86	65.93	48.23	49.30	61.64	63.37
Mean GCSE	4.81	4.90	6.28	6.42	4.88	5.01	6.28	6.49	4.72	4.86	6.14	6.31
No. Items	76	122	78	123	73	93	75	90	78	115	82	125
Mean facility	0.69	0.55	0.64	0.59	0.69	0.52	0.66	0.57	0.67	0.40	0.59	0.49
Mean Discrimination	0.32	0.31	0.35	0.37	0.29	0.33	0.37	0.40	0.33	0.27	0.36	0.40

Table A.2. Descriptive statistics for listening assessments.

Listening	French				German				Spanish			
	Foundation		Higher		Foundation		Higher		Foundation		Higher	
	2017	2018	2017	2018	2017	2018	2017	2018	2017	2018	2017	2018
No. Students	55,174	52,191	54,553	60,462	13,083	15,403	24,675	22,424	31,521	37,027	44,500	46,900
Mean KS2	50.54	49.02	64.15	64.13	51.36	50.46	64.49	66.07	49.53	49.31	62.23	63.43
Mean GCSE	5.01	4.90	6.39	6.42	5.02	4.99	6.33	6.51	4.86	4.86	6.19	6.31
No. Items	75	103	73	113	73	84	75	77	95	98	104	103
Mean facility	0.65	0.45	0.61	0.46	0.68	0.46	0.62	0.59	0.66	0.49	0.66	0.53
Mean Discrimination	0.28	0.25	0.37	0.36	0.23	0.29	0.32	0.78	0.29	0.29	0.33	0.39

Appendix B – Descriptive statistics of item variables

German Reading

Table B.1 Descriptive statistics of item variables in 2017 and 2018 German reading assessments.

	Foundation			Higher		
	2017	2018		2017	2018	
Variable	Mean(SE)	Mean(SE)	T-test	Mean(SE)	Mean(SE)	T-test
Discrimination	0.29(0.01)	0.33(0.02)	1.67	0.37(0.02)	0.4(0.02)	1.19
Facility	**0.69(0.03)**	**0.51(0.03)**	**-4.04*** **	**0.66(0.03)**	**-0.57(0.02)**	**2.69** **
S1	**1.79(0.07)**	**2.15(0.06)**	**4.17*** **	3.08(0.09)	3.28(0.06)	1.88
S3	**1.88(0.07)**	**2.24(0.05)**	**4.27*** **	3.05(0.09)	3.26(0.06)	1.95
S6	**1.97(0.08)**	**2.22(0.06)**	**2.66** **	3.29(0.08)	3.36(0.07)	0.65
No. Words	70.32(6.55)	76.99(2.75)	1.01	**166.25(9.45)**	**-128.43(3.93)**	**3.93*** **
Words per Sentence	**7.68(0.27)**	**9.79(0.3)**	**5.12*** **	11.84(0.36)	-11.52(0.27)	0.74
Lexical Variety	0.83(0.01)	0.81(0.01)	-1.25	**0.72(0.01)**	**0.75(0.01)**	**2.52***
Lexical Density	0.55(0.01)	0.52(0.01)	-1.93	0.52(0)	-0.51(0.01)	1.38
Lexical Familiarity	0.26(0.01)	0.25(0.01)	-0.64	0.29(0.01)	0.3(0.01)	0.45

	Proportion	Proportion	Chi-Squared	Proportion	Proportion	Chi-Squared
Language - French	**0.00**	**0.31**	**25.46*** **	**0.00**	**0.34**	**29.59*** **
Picture included	**0.19**	**0.02**	**11.73*** **	**0.07**	**0.00**	**4.13***
Topic						
Extract	**0.00**	**0.18**	**12.94*** **	**0.00**	**0.20**	**4.13***
Holidays	0.10	0.17	1.40	**0.11**	**0.06**	**14.8*** **
Home and environment	0.12	0.10	0.08	0.07	0.10	0.85
Leisure	0.19	0.10	2.35	0.20	0.16	0.23
Lifestyle	0.41	0.29	2.13	0.35	0.31	0.29
Work and Education	0.18	0.16	0.01	0.28	0.18	0.10
Item Type						
blanks	**0.00**	**0.11**	**6.56***	0.01	0.00	1.90
choose	0.03	0.03	0.00	0.01	0.03	0.01
match	**0.47**	**0.12**	**23.26*** **	0.28	0.19	0.10
MCQ	**0.05**	**0.20**	**6.46***	0.16	0.16	1.44
names	**0.25**	**0.11**	**4.69***	0.24	0.09	0.00
SA	**0.21**	**0.40**	**6.17***	**0.29**	**0.50**	**5.94***
translation	0.00	0.03	0.92	**0.00**	**0.03**	**6.41***

German Listening

Table B.2 Descriptive statistics of item variables in 2017 and 2018 German listening assessments.

	Foundation			Higher		
	2017	2018		2017	2018	
Variable	Mean(SE)	Mean(SE)	T-test	Mean(SE)	Mean(SE)	T-test
Discrimination	**0.23(0.01)**	**0.29(0.01)**	**3.02****	**0.32(0.01)**	**0.38(0.02)**	**2.53 ***
Facility	**0.68(0.03)**	**0.46(0.03)**	**-5.12*****	0.62(0.03)	0.59(0.03)	-0.65
S1	**1.98(0.08)**	**2.41(0.09)**	**3.67*****	**3.05(0.07)**	**3.33(0.05)**	**3.2 ****
S3	**1.98(0.08)**	**2.42(0.08)**	**3.99*****	**3.01(0.07)**	**3.21(0.05)**	**2.33 ***
S6	**1.97(0.08)**	**2.62(0.08)**	**5.61*****	**3.12(0.08)**	**3.51(0.06)**	**3.97 *****
No. Words	**26.33(2.43)**	**43.21(2.73)**	**4.56*****	56.72(3)	54.8(3.11)	-0.44
Words per Sentence	**6.86(0.31)**	**9.44(0.27)**	**6.34*****	**9.09(0.42)**	**11.43(0.39)**	**4.06 *****
Lexical Variety	**0.92(0.01)**	**0.87(0.01)**	**-3.28****	0.84(0.01)	0.85(0.01)	0.67
Lexical Density	0.5(0.01)	0.5(0.01)	0.52	0.49(0.01)	0.49(0.01)	0.14
Lexical Familiarity	0.22(0.02)	0.24(0.01)	1.21	0.26(0.01)	0.25(0.01)	-0.81
Words per second	1.69(0.03)	1.73(0.03)	0.82	1.68(0.03)	1.74(0.03)	1.30
Track Length	**17.15(1.82)**	**26.11(1.77)**	**3.52*****	35.11(1.95)	33.07(2.07)	-0.72
Gap between repeats	12.69(0.6)	13.46(0.55)	0.94	14.91(0.5)	15.88(0.54)	1.31
Time til next track	29.06(2.06)	29.81(2.1)	0.26	34.04(1.67)	35.03(2.48)	0.33
	Proportion	Proportion	Chi-Squared	Proportion	Proportion	Chi-Squared
Language - French	**0.00**	**0.23**	**16.72*****	**0.00**	**0.25**	**18.95*****
Picture included	0.19	0.11	1.61	0.03	0.01	0.00
Topic						
Holidays	0.07	0.10	0.10	0.12	0.14	0.03
Home and environment	0.12	0.21	1.68	0.13	0.14	0.00
Leisure	**0.30**	**0.06**	**14.39*****	**0.28**	**0.12**	**5.39***
Lifestyle	0.25	0.32	0.74	0.20	0.34	2.99
Work and Education	0.26	0.31	0.25	0.27	0.26	0.00
Item Type						
blanks	**0.00**	**0.10**	**5.49***	0.00	0.06	3.20
choose	0.12	0.07	0.69	0.05	0.10	0.73
match	**0.26**	**0.12**	**4.28***	0.16	0.21	0.30
MCQ	0.29	0.32	0.08	0.36	0.21	3.62
names	**0.21**	**0.04**	**9.48****	**0.20**	**0.04**	**7.96****
SA	**0.12**	**0.36**	**10.22****	0.23	0.38	3.37
Gender						
Both	0.37	0.35	0.02	**0.55**	**0.29**	**9.61****
Female	0.34	0.30	0.18	0.23	0.32	1.37
Male	0.29	0.36	0.57	**0.23**	**0.39**	**3.99 ***

Spanish Reading

Table B.3 Descriptive statistics of item variables in 2017 and 2018 Spanish reading assessments.

	Foundation			Higher		
	2017	2018		2017	2018	
Variable	Mean(SE)	Mean(SE)	T-test	Mean(SE)	Mean(SE)	T-test
Discrimination	**0.33(0.01)**	**0.27(0.01)**	**-2.92 ****	**0.36(0.01)**	**0.4(0.01)**	**2.17 ***
Facility	**0.67(0.03)**	**0.4(0.03)**	**-7.40 *****	**0.59(0.03)**	**0.49(0.02)**	**-2.88 ****
S1	**1.61(0.05)**	**2.29(0.05)**	**8.94 *****	**2.39(0.06)**	**2.83(0.05)**	**5.34 *****
S3	**1.53(0.06)**	**2.24(0.06)**	**8.17 *****	**2.33(0.07)**	**2.85(0.05)**	**6.45 *****
S6	**1.69(0.07)**	**2.32(0.06)**	**6.46 *****	**2.66(0.07)**	**2.86(0.06)**	**2.10 ***
No. Words	**48.34(4.18)**	**76.38(3.03)**	**5.56 *****	123.7(11)	119.67(4.11)	-0.39
Words per Sentence	**8.61(0.46)**	**12.64(0.39)**	**6.64 *****	**12.52(0.72)**	**14.79(0.42)**	**2.90 ****
Lexical Variety	**0.82(0.02)**	**0.77(0.01)**	**-3.21 ****	0.73(0.02)	0.72(0.01)	-0.91
Lexical Density	0.54(0.01)	0.53(0.01)	-0.19	**0.51(0.01)**	**0.52(0)**	**2.12 ***
Lexical Familiarity	**0.24(0.01)**	**0.29(0.01)**	**4.10 *****	**0.25(0.01)**	**0.3(0.01)**	**4.16 *****
	Proportion	Proportion	Chi-Squared	Proportion	Proportion	Chi-Squared
Language - French	**0.00**	**0.34**	**31.08*****	**0.00**	**0.28**	**25.68 *****
Picture included	**0.31**	**0.00**	**37.63*****	**0.21**	**0.02**	**17.02 *****
Topic						
Extract	**0.00**	**0.20**	**15.86*****	**0.00**	**0.23**	**17.02 *****
Holidays	0.12	0.14	0.07	**0.16**	**0.20**	**20.24 *****
Home and environment	0.13	0.10	0.23	0.04	0.06	0.33
Leisure	0.13	0.20	1.22	0.16	0.18	0.30
Lifestyle	0.35	0.27	0.96	0.40	0.22	0.02
Work and Education	**0.28**	**0.10**	**10.11****	**0.24**	**0.10**	**6.75 ****
Item Type						
blanks	0.06	0.10	0.26	**0.07**	**0.01**	**6.23 ***
choose	0.01	0.02	0.00	**0.02**	**0.05**	**4.60 ***
match	**0.55**	**0.13**	**37.18*****	0.30	0.12	0.24
MCQ	0.15	0.24	1.76	**0.24**	**0.14**	**9.7 ****
names	0.05	0.11	1.51	0.01	0.04	3.23
SA	**0.17**	**0.37**	**8.71****	0.34	0.62	0.55
translation	0.00	0.03	0.71	**0.00**	**0.02**	**14.71 *****

Spanish Listening

Table B.4 Descriptive statistics of item variables in 2017 and 2018 Spanish listening assessments.

	Foundation			Higher		
	2017	2018		2017	2018	
Variable	Mean(SE)	Mean(SE)	T-test	Mean(SE)	Mean(SE)	T-test
Discrimination	0.29(0.01)	0.29(0.01)	0.34	**0.33(0.01)**	**0.39(0.02)**	**3.03 ****
Facility	**0.66(0.02)**	**0.49(0.03)**	**-4.62 *****	**0.66(0.03)**	**0.53(0.02)**	**-3.74 *****
S1	**1.65(0.05)**	**2.3(0.06)**	**7.97 *****	**2.29(0.06)**	**3.05(0.06)**	**9.26 *****
S3	**1.62(0.05)**	**2.26(0.06)**	**7.83 *****	**2.21(0.06)**	**2.88(0.06)**	**8.34 *****
S6	**1.65(0.05)**	**2.55(0.07)**	**10.26 *****	**2.31(0.06)**	**3.26(0.06)**	**10.96 *****
No. Words	**25.73(2.23)**	**44.19(2.36)**	**5.68 *****	51.42(2.79)	53.53(2.49)	0.56
Words per Sentence	**7.83(0.34)**	**10.34(0.44)**	**4.49 *****	11.55(0.43)	12.78(0.5)	1.85
Lexical Variety	**0.92(0.01)**	**0.86(0.01)**	**-5.56 *****	**0.85(0.01)**	**0.83(0.01)**	**-2.21 ***
Lexical Density	0.54(0.01)	0.53(0.01)	-0.76	0.52(0.01)	0.52(0.01)	0.64
Lexical Familiarity	0.24(0.01)	0.24(0.01)	0.01	0.27(0.01)	0.27(0.01)	0.34
Words per second	1.5(0.04)	1.48(0.03)	-0.52	**1.4(0.02)**	**1.51(0.03)**	**3.01 ****
Track Length	**18.18(1.62)**	**31.17(1.7)**	**5.51 *****	36.6(1.75)	36.6(1.69)	0.00
Gap between repeats	14.19(0.87)	15.6(0.72)	1.25	18.05(0.8)	18.57(0.67)	0.50
Time till next track	32.67(2.24)	38.3(2.11)	1.83	41.32(1.79)	41.6(1.95)	0.10

	Proportion	Proportion	Chi-Squared	Proportion	Proportion	Chi-Squared
Language - French	**0.00**	**0.22**	**21.90 *****	**0.00**	**0.24**	**26.47*****
Picture included	**0.44**	**0.04**	**40.61 *****	**0.25**	**0.00**	**27.22*****
Topic						
Holidays	0.13	0.16	0.27	**0.08**	**0.19**	**5.12***
Home and environment	0.20	0.12	1.62	0.14	0.21	1.26
Leisure	**0.26**	**0.13**	**4.40 ***	**0.26**	**0.06**	**14.19****
Lifestyle	0.24	0.27	0.04	0.22	0.25	0.13
Work and Education	**0.17**	**0.32**	**4.95 ***	0.30	0.28	0.01
Item Type						
blanks	**0.00**	**0.08**	**6.17 ***	0.00	0.05	3.32
choose	**0.02**	**0.13**	**6.90 ****	0.05	0.06	0.00
match	0.26	0.16	2.31	0.15	0.12	0.34
MCQ	**0.29**	**0.13**	**6.64 ****	0.30	0.25	0.34
names	**0.17**	**0.05**	**5.70 ***	**0.12**	**0.03**	**4.52***
SA	**0.25**	**0.44**	**6.58 ***	0.38	0.50	2.14
Gender						
Both	0.14	0.21	1.50	**0.15**	**0.33**	**7.84****
Female	0.43	0.32	2.27	**0.39**	**0.20**	**8.05****
Male	0.43	0.47	0.15	0.45	0.47	0.00

Appendix C – Facility model results

French Reading

Table C.1 Results of facility modelling of French reading assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	-6.27 (25.45)	-0.25	0.81	(21.36)	-1.30	0.19	27.9 (26.16)	1.07	0.29	32.01 (27.45)	1.17	0.24
Mean GCSE	1.47 (5.3)	0.28	0.78	6.33 (4.45)	1.42	0.15	-4.35 (4.15)	-1.05	0.29	-4.3 (4.3)	-1.00	0.32
Year [2017] - 2018	-0.42 (0.51)	-0.82	0.42	0.46 (0.42)	1.08	0.28	0.34 (0.57)	0.60	0.55	0.64 (0.6)	1.06	0.29
Board [AQA] - Pearson	-0.06 (0.18)	-0.35	0.73	-0.11 (0.16)	-0.70	0.49	-0.03 (0.33)	-0.08	0.93	-0.11 (0.36)	-0.32	0.75
Board [AQA] - WJEC	-0.39 (0.25)	-1.59	0.11	-0.57 (0.23)	-2.51	<0.05	0.13 (0.19)	0.69	0.49	0.32 (0.22)	1.49	0.14
language [English] - target	-0.47 (0.21)	-2.27	<0.05	-0.77 (0.21)	-3.70	<0.001	-0.15 (0.18)	-0.84	0.40	-0.41 (0.18)	-2.27	<0.05
S1 (vocab difficulty)				-0.45 (0.16)	-2.86	<0.01				-0.34 (0.16)	-2.12	<0.05
S3 (grammar difficulty)				-0.16 (0.17)	-0.92	0.36				-0.2 (0.16)	-1.28	0.20
S6 (w ork required)				-0.61 (0.17)	-3.48	<0.001				-0.16 (0.14)	-1.11	0.27
Word Count				0 (0)	-0.04	0.97				-0.01 (0)	-2.68	<0.001
Words per sentence				0 (0.02)	0.07	0.95				-0.02 (0.02)	-0.99	0.32
Lexical variety				-1.98 (0.95)	-2.08	<0.05				-4.19 (1.61)	-2.61	<0.001
Lexical density				0.69 (0.86)	0.81	0.42				0.27 (1.73)	0.16	0.88
Lexical unfamiliarity				0.63 (0.78)	0.81	0.42				2.68 (1.41)	1.90	0.06
Pictures included				0.52 (0.21)	2.44	<0.05				0.02 (0.26)	0.08	0.94
Topic [Extract] - Home and environment				-0.13 (0.33)	-0.39	0.70				0.28 (0.38)	0.74	0.46
Topic [Extract] - Home and environment				0.62 (0.25)	2.45	<0.05				0.6 (0.24)	2.56	<0.05
Topic [Extract] - Leisure				0.53 (0.25)	2.09	<0.05				0.83 (0.37)	2.23	<0.05
Topic [Extract] - Lifestyle				0.65 (0.23)	2.78	<0.01				0.66 (0.26)	2.58	<0.05
Topic [Extract] - Work and education				0.47 (0.23)	2.07	<0.05				0.32 (0.24)	1.36	0.17
Item type [blanks] -choose				1.51 (0.56)	2.71	<0.01				1.04 (0.59)	1.76	0.08
Item type [blanks] - match				0.88 (0.35)	2.50	<0.05				0.52 (0.47)	1.11	0.27
Item type [blanks] - MCQ				0.07 (0.34)	0.19	0.85				-0.03 (0.49)	-0.05	0.96
Item type [blanks] - names				0.37 (0.35)	1.04	0.30				0.52 (0.48)	1.07	0.28
Item type [blanks] - SA				-0.68 (0.32)	-2.10	<0.05				-0.28 (0.46)	-0.62	0.54
Item type [blanks] - Translation				-0.7 (0.6)	-1.18	0.24				-0.28 (0.64)	-0.44	0.66
(phi)	2.53 (0.22)	11.34	<0.001	6.59 (0.64)	10.29	<0.001	3.5 (0.31)	11.15	<0.001	5.99 (0.57)	10.60	<0.001
Pseudo - R-squared	0.081			0.614			0.054			0.440		

French Listening

Table C.2 Results of facility modelling of French listening assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	8.99 (25)	0.36	0.72	-35.52 (21.7)	-1.64	0.10	409.45 (153.2)	2.67	<0.01	305 (144.52)	2.11	<0.05
Mean GCSE	-1.52 (4.99)	-0.31	0.76	7.45 (4.37)	1.70	0.09	-63.67 (23.93)	-2.66	<0.01	-46.58 (22.55)	-2.07	<0.05
Year [2017] - 2018	-1.11 (0.56)	-2.00	<0.05	0.4 (0.49)	0.82	0.41	1.17 (0.68)	1.73	0.08	0.9 (0.65)	1.37	0.17
Board [AQA] - Pearson	0.16 (0.25)	0.63	0.53	0.33 (0.26)	1.30	0.19	-3.13 (1.32)	-2.37	<0.05	-2.61 (1.24)	-2.10	<0.05
Board [AQA] - WJEC	1.79 (0.36)	4.92	<0.001	0.63 (0.36)	1.72	0.09	0.56 (0.41)	1.36	0.17	0.08 (0.46)	0.17	0.86
language [English] - target	0.3 (0.28)	1.09	0.27	0.41 (0.28)	1.48	0.14	0.19 (0.21)	0.92	0.36	0.14 (0.24)	0.57	0.57
Speech Speed (sec)	0.15 (0.18)	0.85	0.39	-0.04 (0.18)	-0.20	0.85	-0.34 (0.35)	-0.98	0.33	-1.42 (0.54)	-2.61	<0.001
Pause length (sec)	-0.09 (0.03)	-3.53	<0.001	0.01 (0.02)	0.34	0.73	-0.06 (0.02)	-2.38	<0.05	-0.03 (0.03)	-1.31	0.19
Time betw een tracks (sec)	-0.01 (0.01)	-1.18	0.24	0 (0)	0.06	0.95	0 (0.01)	-0.85	0.40	0 (0.01)	-0.42	0.67
S1 (vocab difficulty)				-0.09 (0.22)	-0.40	0.69				-0.48 (0.18)	-2.67	<0.001
S3 (grammar difficulty)				-0.45 (0.21)	-2.16	<0.05				0.17 (0.19)	0.90	0.37
S6 (w ork required)				-0.08 (0.19)	-0.40	0.69				-0.18 (0.18)	-1.01	0.31
Word count				0.01 (0.02)	0.63	0.53				0.03 (0.02)	1.86	0.06
Words per sentence				-0.01 (0.02)	-0.38	0.70				0 (0.02)	0.20	0.84
Lexical variety				0.56 (1.01)	0.55	0.58				-1.24 (1.44)	-0.86	0.39
lexical density				0.74 (0.73)	1.01	0.31				-1.3 (1.4)	-0.92	0.36
Lexical unfamiliarity				-0.5 (0.78)	-0.63	0.53				0.49 (1.03)	0.48	0.63
Pictures included				0.58 (0.22)	2.67	<0.01				-0.06 (0.32)	-0.19	0.85
Topic [Holidays] - Home and environment				0.53 (0.25)	2.13	<0.05				0.32 (0.32)	0.99	0.32
Topic [Holidays] - Leisure				0.53 (0.24)	2.20	<0.05				0.6 (0.33)	1.82	0.07
Topic [Holidays] - Lifestyle				0.27 (0.22)	1.23	0.22				0.78 (0.33)	2.40	<0.05
Topic [Holidays] - Work and education				0.55 (0.2)	2.72	<0.01				0.72 (0.26)	2.75	<0.001
Item type [blanks] -choose				0.23 (0.49)	0.48	0.63				-0.06 (0.59)	-0.10	0.92
Item type [blanks] - match				-0.45 (0.44)	-1.00	0.32				-0.11 (0.6)	-0.19	0.85
Item type [blanks] - MCQ				-0.35 (0.44)	-0.80	0.42				-0.18 (0.54)	-0.34	0.74
Item type [blanks] - names				-0.9 (0.47)	-1.92	0.05				-1.09 (0.58)	-1.86	0.06
Item type [blanks] - SA				-1.64 (0.42)	-3.94	<0.001				-0.97 (0.54)	-1.78	0.07
Track length (sec)				-0.03 (0.02)	-1.55	0.12				-0.06 (0.03)	-2.09	<0.05
Speaker gender [both] - Female				-0.69 (0.23)	-3.02	<0.01				-0.27 (0.22)	-1.24	0.22
Speaker gender [both] - Male				-0.45 (0.23)	-1.97	<0.05				-0.1 (0.2)	-0.48	0.63
(phi)	3.12 (0.3)	10.53	<0.001	7.38 (0.75)	9.78	<0.001	3.75 (0.35)	10.68	<0.001	5.82 (0.57)	10.25	<0.001
Pseudo - R-squared	0.237			0.664			0.188			0.470		

German Reading

Table C.3 Results of facility modelling of German reading assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	-1.59 (5.07)	-0.31	0.75	-3.44 (4.55)	-0.76	0.45	14.95 (13.39)	1.12	0.26	-6.16 (11.4)	-0.54	0.59
Mean GCSE	0.39 (1.04)	0.38	0.71	0.71 (0.89)	0.80	0.43	-2.29 (2.13)	-1.07	0.28	1.4 (1.77)	0.79	0.43
Year [2017] - 2018	-0.56 (0.23)	-2.43	<0.05	-0.04 (0.24)	-0.15	0.88	0.02 (0.49)	0.04	0.97	-0.47 (0.42)	-1.12	0.26
Board [AQA] - Pearson	0.67 (0.17)	3.88	<0.001	0.8 (0.18)	4.55	<0.001	-0.01 (0.2)	-0.04	0.97	0.08 (0.19)	0.43	0.66
Board [AQA] - WJEC	0.47 (0.29)	1.61	0.11	0.25 (0.29)	0.84	0.40	0.44 (0.3)	1.46	0.15	0.14 (0.29)	0.47	0.64
language [English] - target	-0.53 (0.23)	-2.28	<0.05	-0.36 (0.24)	-1.50	0.13	0.05 (0.2)	0.24	0.81	-0.05 (0.19)	-0.26	0.80
S1 (vocab difficulty)				-1.21 (0.28)	-4.27	<0.001				-0.51 (0.25)	-2.09	<0.05
S3 (grammar difficulty)				-0.4 (0.27)	-1.51	0.13				0.18 (0.26)	0.69	0.49
S6 (w ork required)				-0.09 (0.28)	-0.32	0.75				-0.56 (0.23)	-2.41	<0.05
Word Count				0.01 (0)	2.18	<0.05				0 (0)	0.54	0.59
Words per sentence				0.04 (0.04)	1.01	0.31				0.05 (0.03)	1.71	0.09
Lexical variety				0.89 (1.49)	0.60	0.55				-0.75 (2.36)	-0.32	0.75
Lexical density				1.73 (1.08)	1.60	0.11				-0.98 (1.92)	-0.51	0.61
Lexical unfamiliarity				-0.01 (0.91)	-0.02	0.99				0.99 (1.42)	0.70	0.49
Pictures included				0.48 (0.38)	1.28	0.20				-0.51 (0.43)	-1.19	0.23
Topic [Extract] - Home and environment				0.5 (0.32)	1.57	0.12				-0.53 (0.3)	-1.78	0.07
Topic [Extract] - Home and environment				0.63 (0.32)	1.94	0.05				0.36 (0.31)	1.18	0.24
Topic [Extract] - Leisure				0.53 (0.33)	1.58	0.12				0.26 (0.26)	0.97	0.33
Topic [Extract] - Lifestyle				0.38 (0.27)	1.42	0.16				0.35 (0.24)	1.45	0.15
Topic [Extract] - Work and education				0.44 (0.29)	1.52	0.13				0.18 (0.27)	0.66	0.51
Item type [blanks] -choose				1.93 (0.54)	3.61	<0.001				1.71 (0.83)	2.05	<0.05
Item type [blanks] - match				0.35 (0.35)	0.99	0.32				1.09 (0.79)	1.38	0.17
Item type [blanks] - MCQ				0.56 (0.39)	1.44	0.15				0.97 (0.8)	1.21	0.23
Item type [blanks] - names				0.6 (0.4)	1.52	0.13				1.03 (0.77)	1.34	0.18
Item type [blanks] - SA				0.04 (0.36)	0.10	0.92				0.21 (0.78)	0.27	0.79
Item type [blanks] - Translation				1.07 (0.62)	1.72	0.08				0.73 (0.89)	0.82	0.41
(phi)	2.53 (0.24)	10.31	<0.001	5.61 (0.59)	9.46	<0.001	3.69 (0.37)	10.05	<0.001	7.65 (0.81)	9.48	<0.001
Pseudo - R-squared	0.190			0.604			0.075			0.542		

German Listening

Table C.4 Results of facility modelling of German listening assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	-2.08 (4.5)	-0.46	0.64	0.22 (4.56)	0.05	0.96	-20.78 (12.45)	-1.67	0.10	0.38 (13.98)	0.03	0.98
Mean GCSE	0.83 (0.91)	0.91	0.36	-0.22 (0.82)	-0.27	0.79	3.74 (1.98)	1.89	0.06	1.04 (2.21)	0.47	0.64
Year [2017] - 2018	-0.83 (0.18)	-4.60	<0.001	0.03 (0.18)	0.17	0.86	-0.58 (0.34)	-1.68	0.09	0.1 (0.4)	0.25	0.80
Board [AQA] - Pearson	-0.8 (0.32)	-2.45	<0.05	-0.63 (0.35)	-1.78	0.08	-0.24 (0.32)	-0.74	0.46	0.36 (0.43)	0.84	0.40
Board [AQA] - WJEC	-0.1 (0.35)	-0.29	0.77	-0.69 (0.35)	-2.00	<0.05	-0.25 (0.36)	-0.70	0.48	-0.02 (0.36)	-0.06	0.95
language [English] - target	0.03 (0.27)	0.10	0.92	-0.33 (0.27)	-1.20	0.23	-0.09 (0.27)	-0.31	0.75	0.04 (0.3)	0.12	0.90
Speech Speed (sec)	-0.14 (0.36)	-0.38	0.70	-0.35 (0.39)	-0.89	0.38	-0.52 (0.36)	-1.45	0.15	-0.14 (0.51)	-0.28	0.78
Pause length (sec)	-0.03 (0.03)	-1.35	0.18	0.03 (0.03)	1.09	0.28	-0.08 (0.02)	-3.36	<0.001	0 (0.03)	0.04	0.96
Time betw een tracks (sec)	-0.01 (0.01)	-1.76	0.08	0.01 (0.01)	1.86	0.06	-0.01 (0.01)	-0.83	0.41	-0.02 (0.01)	-1.99	<0.05
S1 (vocab difficulty)				-0.35 (0.3)	-1.18	0.24				-0.76 (0.35)	-2.17	<0.05
S3 (grammar difficulty)				-0.21 (0.31)	-0.68	0.49				0.18 (0.33)	0.54	0.59
S6 (w ork required)				-0.54 (0.27)	-2.00	<0.05				-0.3 (0.29)	-1.02	0.31
Word count				0.04 (0.02)	1.95	0.05				-0.03 (0.02)	-1.22	0.22
Words per sentence				0 (0.04)	-0.07	0.95				0.04 (0.02)	1.63	0.10
Lexical variety				2.47 (1.52)	1.63	0.10				-3.53 (1.75)	-2.02	<0.05
lexical density				0.33 (1.03)	0.32	0.75				0.09 (1.32)	0.07	0.94
Lexical unfamiliarity				-0.71 (0.72)	-0.98	0.33				-1.77 (1.24)	-1.43	0.15
Pictures included				0.3 (0.25)	1.20	0.23				-0.81 (0.7)	-1.17	0.24
Topic [Holidays] - Home and environment				-0.16 (0.33)	-0.47	0.64				-0.21 (0.33)	-0.64	0.52
Topic [Holidays] - Leisure				0.64 (0.37)	1.76	0.08				-0.07 (0.33)	-0.22	0.83
Topic [Holidays] - Lifestyle				0.38 (0.38)	1.01	0.31				-0.03 (0.31)	-0.09	0.93
Topic [Holidays] - Work and education				0.55 (0.32)	1.75	0.08				-0.2 (0.27)	-0.71	0.48
Item type [blanks] -choose				1.39 (0.47)	2.94	<0.01				1.84 (0.6)	3.05	<0.001
Item type [blanks] - match				1.07 (0.41)	2.59	<0.01				1.02 (0.55)	1.87	0.06
Item type [blanks] - MCQ				1.06 (0.44)	2.43	<0.05				0.9 (0.58)	1.55	0.12
Item type [blanks] - names				0.93 (0.44)	2.12	<0.05				0.25 (0.59)	0.42	0.68
Item type [blanks] - SA				-0.5 (0.41)	-1.24	0.22				-0.45 (0.59)	-0.76	0.45
Track length (sec)				-0.05 (0.03)	-1.43	0.15				0.03 (0.03)	0.86	0.39
Speaker gender [both] - Female				0.38 (0.24)	1.57	0.12				-0.25 (0.22)	-1.11	0.27
Speaker gender [both] - Male				0.34 (0.25)	1.40	0.16				-0.01 (0.19)	-0.04	0.97
(phi)	2.73 (0.27)	9.94	<0.001	7.01 (0.77)	9.14	<0.001	3.6 (0.37)	9.64	<0.001	6.56 (0.72)	9.16	<0.001
Pseudo - R-squared	0.236			0.685			0.128			0.511		

Spanish Reading

Table C.5 Results of facility modelling of Spanish reading assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	-15.67 (8.42)	-1.86	0.06	1.62 (7.28)	0.22	0.82	-31.19 (21.86)	-1.43	0.15	5.01 (18.43)	0.27	0.79
Mean GCSE	3.43 (1.77)	1.93	0.05	1.02 (1.56)	0.65	0.52	5.09 (3.54)	1.44	0.15	-0.36 (2.97)	-0.12	0.90
Year [2017] - 2018	-1.46 (0.27)	-5.46	<0.001	-0.38 (0.25)	-1.49	0.14	-1.18 (0.56)	-2.13	<0.05	-0.43 (0.49)	-0.89	0.37
Board [AQA] - Pearson	0.22 (0.17)	1.31	0.19	0.36 (0.16)	2.29	<0.05	0.83 (0.44)	1.87	0.06	-0.06 (0.39)	-0.15	0.88
Board [AQA] - WJEC	0.32 (0.37)	0.85	0.39	1.07 (0.32)	3.37	<0.001	0.14 (0.33)	0.41	0.68	0.23 (0.29)	0.79	0.43
language [English] - target	0.12 (0.2)	0.61	0.54	0.06 (0.19)	0.32	0.75	0.27 (0.19)	1.44	0.15	0.04 (0.17)	0.23	0.82
S1 (vocab difficulty)				-0.28 (0.24)	-1.16	0.25				-0.3 (0.2)	-1.52	0.13
S3 (grammar difficulty)				-0.06 (0.2)	-0.29	0.77				0.21 (0.18)	1.15	0.25
S6 (w ork required)				-0.65 (0.2)	-3.30	<0.001				-0.68 (0.17)	-4.08	<0.001
Word Count				-0.01 (0)	-2.54	<0.05				0 (0)	-2.68	<0.001
Words per sentence				-0.01 (0.02)	-0.25	0.80				0 (0.02)	0.26	0.80
Lexical variety				-5.55 (1.28)	-4.33	<0.001				-3.28 (1.1)	-2.98	<0.001
Lexical density				0.23 (0.85)	0.27	0.79				2.09 (1.35)	1.56	0.12
Lexical unfamiliarity				1.12 (0.82)	1.37	0.17				0.95 (0.92)	1.03	0.30
Pictures included				0.56 (0.24)	2.29	<0.05				-0.78 (0.25)	-3.13	<0.001
Topic [Extract] - Home and environment				-0.26 (0.28)	-0.91	0.36				-0.13 (0.2)	-0.64	0.52
Topic [Extract] - Home and environment				0.47 (0.3)	1.56	0.12				0.32 (0.29)	1.10	0.27
Topic [Extract] - Leisure				0.32 (0.28)	1.16	0.24				-0.05 (0.2)	-0.26	0.80
Topic [Extract] - Lifestyle				0.07 (0.27)	0.27	0.79				-0.13 (0.2)	-0.68	0.50
Topic [Extract] - Work and education				0.39 (0.29)	1.34	0.18				0.51 (0.24)	2.12	<0.05
Item type [blanks] -choose				0.75 (0.53)	1.41	0.16				1.75 (0.45)	3.89	<0.001
Item type [blanks] - match				0.4 (0.24)	1.68	0.09				1.93 (0.38)	5.01	<0.001
Item type [blanks] - MCQ				0.08 (0.24)	0.32	0.75				1.74 (0.4)	4.33	<0.001
Item type [blanks] - names				0.4 (0.32)	1.26	0.21				2.11 (0.5)	4.26	<0.001
Item type [blanks] - SA				-1.11 (0.25)	-4.48	<0.001				0.97 (0.37)	2.61	<0.001
Item type [blanks] - Translation				0.38 (0.51)	0.74	0.46				1.18 (0.53)	2.23	<0.05
(phi)	2.59 (0.23)	11.12	<0.001	6.75 (0.67)	10.14	<0.001	3.58 (0.32)	11.36	<0.001	7.55 (0.71)	10.66	<0.001
Pseudo - R-squared	0.226			0.647			0.109			0.587		

Spanish Listening

Table C.6 Results of facility modelling of Spanish listening assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	-2.52 (3.31)	-0.76	0.45	2.85 (3.44)	0.83	0.41	-17.14 (7.66)	-2.24	<0.05	3.71 (7.45)	0.50	0.62
Mean GCSE	0.77 (0.69)	1.11	0.27	-0.01 (0.61)	-0.01	0.99	2.91 (1.25)	2.32	<0.05	-0.05 (1.18)	-0.05	0.96
Year [2017] - 2018	-0.62 (0.18)	-3.55	<0.001	0.15 (0.2)	0.75	0.45	-0.65 (0.15)	-4.25	<0.001	0.12 (0.16)	0.76	0.45
Board [AQA] - Pearson	0.08 (0.22)	0.36	0.72	-0.37 (0.24)	-1.53	0.13	0.93 (0.27)	3.44	<0.001	0.07 (0.35)	0.21	0.84
Board [AQA] - WJEC	0.5 (0.28)	1.76	0.08	0.34 (0.25)	1.36	0.17	0.96 (0.23)	4.24	<0.001	0.88 (0.25)	3.52	<0.001
language [English] - target	-0.14 (0.26)	-0.52	0.60	-0.43 (0.29)	-1.49	0.14	-0.13 (0.22)	-0.57	0.57	-0.14 (0.2)	-0.69	0.49
Speech Speed (sec)	-0.25 (0.26)	-0.95	0.34	-0.06 (0.35)	-0.19	0.85	-0.14 (0.33)	-0.42	0.67	0.79 (0.44)	1.81	0.07
Pause length (sec)	-0.03 (0.02)	-1.47	0.14	-0.03 (0.02)	-1.64	0.10	-0.03 (0.01)	-2.61	<0.01	-0.01 (0.01)	-0.77	0.44
Time betw een tracks (sec)	0 (0)	-0.06	0.95	0 (0)	-0.36	0.72	0 (0)	-0.35	0.73	-0.01 (0)	-1.93	0.05
S1 (vocab difficulty)				-0.88 (0.23)	-3.89	<0.001				-0.21 (0.19)	-1.12	0.26
S3 (grammar difficulty)				0.05 (0.22)	0.23	0.82				-0.09 (0.18)	-0.53	0.60
S6 (w ork required)				-0.36 (0.18)	-1.97	<0.05				-0.54 (0.14)	-3.72	<0.001
Word count				0 (0.02)	0.26	0.80				-0.02 (0.02)	-1.10	0.27
Words per sentence				-0.05 (0.02)	-2.41	<0.05				-0.03 (0.02)	-1.97	<0.05
Lexical variety				0.96 (1.42)	0.67	0.50				-1.61 (1.32)	-1.23	0.22
lexical density				1.07 (0.77)	1.39	0.16				1.27 (1.01)	1.25	0.21
Lexical unfamiliarity				-0.48 (0.68)	-0.71	0.48				-1.02 (0.79)	-1.29	0.20
Pictures included				-0.1 (0.23)	-0.42	0.68				-0.46 (0.24)	-1.91	0.06
Topic [Holidays] - Home and environment				0.05 (0.22)	0.23	0.82				-0.23 (0.23)	-1.01	0.31
Topic [Holidays] - Leisure				0.03 (0.22)	0.14	0.89				-0.19 (0.22)	-0.86	0.39
Topic [Holidays] - Lifestyle				0.13 (0.21)	0.60	0.55				-0.24 (0.21)	-1.14	0.26
Topic [Holidays] - Work and education				0.15 (0.21)	0.71	0.48				-0.11 (0.21)	-0.54	0.59
Item type [blanks] -choose				0.56 (0.45)	1.25	0.21				0.64 (0.55)	1.17	0.24
Item type [blanks] - match				-0.17 (0.42)	-0.40	0.69				0.17 (0.47)	0.37	0.71
Item type [blanks] - MCQ				-0.38 (0.43)	-0.89	0.38				0.46 (0.44)	1.04	0.30
Item type [blanks] - names				-0.63 (0.51)	-1.24	0.21				0.91 (0.54)	1.68	0.09
Item type [blanks] - SA				-1.27 (0.42)	-3.04	<0.01				-0.63 (0.46)	-1.37	0.17
Track length (sec)				0 (0.02)	0.21	0.84				0.02 (0.02)	1.01	0.31
Speaker gender [both] - Female				-0.03 (0.22)	-0.14	0.89				0.01 (0.19)	0.05	0.96
Speaker gender [both] - Male				-0.11 (0.21)	-0.52	0.60				0.14 (0.17)	0.84	0.40
(phi)	2.84 (0.26)	11.11	<0.001	7.03 (0.69)	10.20	<0.001	3.77 (0.34)	11.15	<0.001	8.36 (0.79)	10.54	<0.001
Pseudo - R-squared	0.140			0.617			0.220			0.641		

Appendix D – Discrimination model results

French Reading

Table D.1 Results of discrimination modelling of French reading assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	-5.03 (2.94)	-1.71	0.09	-6.39 (2.97)	-2.15	<0.05	-9.24 (3.35)	-2.76	<0.01	-10.53 (3.74)	-2.81	<0.001
Mean GCSE	1.12 (0.61)	1.83	0.07	1.43 (0.62)	2.31	<0.05	1.52 (0.53)	2.86	<0.01	1.67 (0.59)	2.84	<0.001
Year [2017] - 2018	-0.14 (0.06)	-2.30	<0.05	-0.14 (0.06)	-2.31	<0.05	-0.2 (0.07)	-2.70	<0.01	-0.26 (0.08)	-3.09	<0.001
Board [AQA] - Pearson	-0.06 (0.02)	-3.15	<0.01	-0.05 (0.02)	-2.49	<0.05	0.11 (0.04)	2.71	<0.01	0.13 (0.05)	2.54	<0.05
Board [AQA] - WJEC	0.05 (0.03)	1.71	0.09	0 (0.03)	-0.12	0.90	0.05 (0.02)	2.26	<0.05	0.05 (0.03)	1.86	0.07
language [English] - target	-0.01 (0.02)	-0.59	0.56	0.02 (0.03)	0.54	0.59	0.01 (0.02)	0.22	0.83	0.04 (0.03)	1.41	0.16
S1 (vocab difficulty)				-0.01 (0.02)	-0.42	0.68				-0.02 (0.02)	-1.02	0.31
S3 (grammar difficulty)				-0.01 (0.02)	-0.58	0.56				0.03 (0.02)	1.28	0.20
S6 (w ork required)				-0.03 (0.02)	-1.43	0.15				0 (0.02)	-0.04	0.97
Word Count				0 (0)	1.95	0.05				0 (0)	-0.62	0.54
Words per sentence				0 (0)	-0.47	0.64				0 (0)	1.09	0.28
Lexical variety				-0.11 (0.13)	-0.86	0.39				0.27 (0.22)	1.25	0.21
Lexical density				-0.01 (0.11)	-0.11	0.91				0.12 (0.23)	0.51	0.61
Lexical unfamiliarity				-0.09 (0.1)	-0.84	0.40				0.08 (0.19)	0.42	0.67
Pictures included				0.03 (0.03)	0.91	0.36				0.01 (0.04)	0.35	0.72
Topic [Extract] - Home and environment				0.05 (0.05)	1.06	0.29				-0.01 (0.05)	-0.24	0.81
Topic [Extract] - Home and environment				0.06 (0.04)	1.59	0.11				-0.05 (0.03)	-1.51	0.13
Topic [Extract] - Leisure				0.06 (0.04)	1.56	0.12				0.04 (0.05)	0.85	0.40
Topic [Extract] - Lifestyle				0.06 (0.03)	1.88	0.06				-0.05 (0.04)	-1.40	0.16
Topic [Extract] - Work and education				0.05 (0.03)	1.44	0.15				-0.02 (0.03)	-0.51	0.61
Item type [blanks] -choose				0.04 (0.07)	0.53	0.60				0.22 (0.08)	2.73	<0.001
Item type [blanks] - match				-0.02 (0.05)	-0.43	0.67				0.02 (0.07)	0.35	0.73
Item type [blanks] - MCQ				-0.08 (0.05)	-1.61	0.11				0.08 (0.07)	1.11	0.27
Item type [blanks] - names				-0.09 (0.05)	-1.88	0.06				0.1 (0.07)	1.41	0.16
Item type [blanks] - SA				0 (0.05)	-0.05	0.96				0.13 (0.06)	1.95	0.05
Item type [blanks] - Translation				0.3 (0.09)	3.51	<0.001				0.35 (0.09)	3.91	<0.001
Pseudo - R-squared	0.152			0.375			0.088			0.340		

French Listening

Table D.1 Results of discrimination modelling of French listening assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	14.85 (3.21)	4.62	<0.001	12.86 (3.48)	3.69	<0.001	63.68 (18.66)	3.41	<0.001	76.96 (19.33)	3.98	<0.001
Mean GCSE	-2.94 (0.64)	-4.58	<0.001	-2.54 (0.7)	-3.61	<0.001	-9.89 (2.91)	-3.39	<0.001	-11.89 (3.02)	-3.94	<0.001
Year [2017] - 2018	-0.34 (0.07)	-4.75	<0.001	-0.26 (0.08)	-3.26	<0.01	0.24 (0.08)	2.92	<0.01	0.34 (0.09)	3.84	<0.001
Board [AQA] - Pearson	0.14 (0.03)	4.15	<0.001	0.19 (0.04)	4.59	<0.001	-0.56 (0.16)	-3.44	<0.001	-0.64 (0.17)	-3.81	<0.001
Board [AQA] - WJEC	0.09 (0.05)	1.90	0.06	0.08 (0.06)	1.37	0.17	-0.06 (0.05)	-1.11	0.27	-0.12 (0.06)	-1.83	0.07
language [English] - target	-0.09 (0.04)	-2.43	<0.05	-0.01 (0.05)	-0.14	0.89	0.04 (0.03)	1.45	0.15	0.03 (0.03)	0.90	0.37
Speech Speed (sec)	0.04 (0.02)	1.86	0.06	0.04 (0.03)	1.28	0.20	-0.02 (0.04)	-0.37	0.71	-0.14 (0.07)	-1.81	0.07
Pause length (sec)	0 (0)	0.81	0.42	0.01 (0)	1.53	0.13	0.01 (0)	1.86	0.06	0.01 (0)	1.96	0.05
Time betw een tracks (sec)	0 (0)	-1.69	0.09	0 (0)	0.25	0.80	0 (0)	-0.37	0.71	0 (0)	-0.67	0.50
S1 (vocab difficulty)				0 (0.04)	0.14	0.89				-0.05 (0.02)	-2.06	<0.05
S3 (grammar difficulty)				0.05 (0.03)	1.42	0.16				0.02 (0.03)	0.86	0.39
S6 (w ork required)				-0.05 (0.03)	-1.63	0.10				0 (0.03)	-0.18	0.86
Word count				0 (0)	-0.11	0.91				0 (0)	1.20	0.23
Words per sentence				-0.01 (0)	-2.04	<0.05				0 (0)	-0.47	0.64
Lexical variety				-0.14 (0.16)	-0.90	0.37				-0.13 (0.19)	-0.65	0.52
lexical density				0.02 (0.12)	0.21	0.83				-0.21 (0.19)	-1.08	0.28
Lexical unfamiliarity				-0.05 (0.13)	-0.42	0.68				-0.03 (0.14)	-0.25	0.80
Pictures included				-0.02 (0.03)	-0.59	0.56				0.08 (0.04)	1.73	0.09
Topic [Holidays] - Home and environment				0.04 (0.04)	1.11	0.27				0 (0.04)	0.03	0.98
Topic [Holidays] - Leisure				0 (0.04)	-0.05	0.96				-0.01 (0.05)	-0.13	0.90
Topic [Holidays] - Lifestyle				-0.07 (0.03)	-2.04	<0.05				0.01 (0.04)	0.27	0.79
Topic [Holidays] - Work and education				0 (0.03)	0.03	0.98				0 (0.04)	0.04	0.97
Item type [blanks] -choose				0.19 (0.08)	2.40	<0.05				0.16 (0.08)	1.89	0.06
Item type [blanks] - match				0.2 (0.07)	2.82	<0.01				0.12 (0.08)	1.47	0.14
Item type [blanks] - MCQ				0.14 (0.07)	2.05	<0.05				0.03 (0.08)	0.39	0.70
Item type [blanks] - names				0.17 (0.08)	2.31	<0.05				0.08 (0.08)	1.04	0.30
Item type [blanks] - SA				0.17 (0.07)	2.60	<0.05				0.11 (0.08)	1.50	0.13
Track length (sec)				0 (0)	-0.88	0.38				-0.01 (0)	-1.63	0.11
Speaker gender [both] - Female				0 (0.04)	-0.12	0.91				-0.01 (0.03)	-0.46	0.64
Speaker gender [both] - Male				-0.02 (0.04)	-0.65	0.51				-0.02 (0.03)	-0.77	0.44
Pseudo - R-squared	0.206			0.390			0.130			0.306		

German Reading

Table D.3 Results of discrimination modelling of German reading assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	1.59 (0.55)	2.92	<0.01	1.2 (0.56)	2.15	<0.05	-1.75 (1.92)	-0.91	0.36	1.31 (1.74)	0.75	0.45
Mean GCSE	-0.27 (0.11)	-2.42	<0.05	-0.2 (0.11)	-1.87	0.06	0.34 (0.31)	1.10	0.27	-0.08 (0.27)	-0.30	0.77
Year [2017] - 2018	0.1 (0.02)	3.90	<0.001	0.08 (0.03)	2.63	<0.01	-0.06 (0.07)	-0.85	0.40	0 (0.06)	0.02	0.98
Board [AQA] - Pearson	0 (0.02)	0.07	0.94	0.02 (0.02)	0.96	0.34	0 (0.03)	-0.09	0.93	-0.05 (0.03)	-1.71	0.09
Board [AQA] - WJEC	0.25 (0.03)	7.98	<0.001	0.23 (0.04)	6.21	<0.001	0.24 (0.04)	5.49	<0.001	0.21 (0.04)	4.82	<0.001
language [English] - target	-0.09 (0.03)	-3.70	<0.001	-0.04 (0.03)	-1.35	0.18	0.01 (0.03)	0.21	0.83	0.01 (0.03)	0.20	0.84
S1 (vocab difficulty)				-0.06 (0.04)	-1.61	0.11				0.08 (0.04)	2.17	<0.05
S3 (grammar difficulty)				0.03 (0.03)	0.83	0.41				-0.05 (0.04)	-1.29	0.20
S6 (w ork required)				0.02 (0.04)	0.47	0.64				0 (0.04)	0.11	0.91
Word Count				0 (0)	0.89	0.37				0 (0)	0.09	0.93
Words per sentence				0 (0.01)	-0.15	0.88				-0.01 (0)	-2.82	<0.001
Lexical variety				-0.05 (0.18)	-0.28	0.78				-0.03 (0.35)	-0.10	0.92
Lexical density				0.17 (0.13)	1.26	0.21				0.24 (0.28)	0.85	0.40
Lexical unfamiliarity				0.07 (0.11)	0.67	0.50				-0.63 (0.21)	-2.95	<0.001
Pictures included				-0.02 (0.05)	-0.39	0.70				-0.03 (0.06)	-0.44	0.66
Topic [Extract] - Home and environment				-0.09 (0.04)	-2.08	<0.05				0.13 (0.05)	2.80	<0.001
Topic [Extract] - Home and environment				0.04 (0.04)	1.08	0.28				-0.12 (0.05)	-2.52	<0.05
Topic [Extract] - Leisure				0.03 (0.04)	0.76	0.45				-0.02 (0.04)	-0.54	0.59
Topic [Extract] - Lifestyle				0 (0.03)	-0.02	0.98				-0.05 (0.04)	-1.39	0.17
Topic [Extract] - Work and education				-0.01 (0.04)	-0.16	0.87				-0.06 (0.04)	-1.53	0.13
Item type [blanks] -choose				-0.01 (0.07)	-0.09	0.93				-0.22 (0.13)	-1.70	0.09
Item type [blanks] - match				-0.04 (0.04)	-0.92	0.36				-0.3 (0.12)	-2.41	<0.05
Item type [blanks] - MCQ				-0.06 (0.05)	-1.31	0.19				-0.31 (0.13)	-2.43	<0.05
Item type [blanks] - names				0 (0.05)	0.06	0.95				-0.29 (0.12)	-2.42	<0.05
Item type [blanks] - SA				0 (0.05)	0.01	1.00				-0.19 (0.12)	-1.58	0.12
Item type [blanks] - Translation				0.22 (0.08)	2.79	<0.01				0.02 (0.14)	0.14	0.89
Pseudo - R-squared	0.381			0.533			0.251			0.561		

German Listening

Table D.4 Results of discrimination modelling of German listening assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	0.93 (0.44)	2.14	<0.05	0.67 (0.5)	1.35	0.18	2.16 (1.49)	1.45	0.15	-2.86 (1.69)	-1.70	0.09
Mean GCSE	-0.14 (0.09)	-1.58	0.12	-0.09 (0.09)	-1.02	0.31	-0.34 (0.24)	-1.42	0.16	0.41 (0.27)	1.53	0.13
Year [2017] - 2018	0.07 (0.02)	3.82	<0.001	0.11 (0.02)	5.50	<0.001	0.09 (0.04)	2.17	<0.05	-0.05 (0.05)	-1.06	0.29
Board [AQA] - Pearson	0.02 (0.03)	0.61	0.54	0.05 (0.04)	1.26	0.21	-0.01 (0.04)	-0.13	0.90	-0.2 (0.05)	-3.96	<0.001
Board [AQA] - WJEC	0.21 (0.03)	6.18	<0.001	0.23 (0.04)	6.02	<0.001	0.19 (0.04)	4.37	<0.001	0.11 (0.04)	2.46	<0.05
language [English] - target	-0.05 (0.03)	-1.79	0.08	-0.01 (0.03)	-0.28	0.78	-0.01 (0.03)	-0.35	0.72	0.03 (0.03)	0.82	0.41
Speech Speed (sec)	-0.01 (0.04)	-0.33	0.74	0.03 (0.04)	0.68	0.50	0.07 (0.04)	1.52	0.13	0.09 (0.06)	1.45	0.15
Pause length (sec)	0 (0)	-0.23	0.82	0 (0)	-0.60	0.55	0.01 (0)	3.04	<0.01	-0.01 (0)	-3.43	<0.001
Time between tracks (sec)	0 (0)	-0.90	0.37	0 (0)	0.96	0.34	0 (0)	1.07	0.28	0 (0)	3.07	<0.001
S1 (vocab difficulty)				-0.07 (0.03)	-2.04	<0.05				0 (0.04)	-0.04	0.97
S3 (grammar difficulty)				0.08 (0.03)	2.29	<0.05				0.05 (0.04)	1.34	0.18
S6 (work required)				0 (0.03)	-0.07	0.95				-0.02 (0.03)	-0.55	0.58
Word count				0 (0)	-0.83	0.41				0 (0)	0.20	0.84
Words per sentence				0.01 (0)	2.20	<0.05				0 (0)	0.04	0.97
Lexical variety				-0.26 (0.17)	-1.52	0.13				0.47 (0.21)	2.26	<0.05
lexical density				0.14 (0.11)	1.25	0.21				-0.13 (0.16)	-0.85	0.40
Lexical unfamiliarity				-0.17 (0.08)	-2.24	<0.05				0.09 (0.15)	0.59	0.56
Pictures included				-0.02 (0.03)	-0.92	0.36				-0.07 (0.08)	-0.89	0.38
Topic [Holidays] - Home and environment				-0.02 (0.04)	-0.59	0.56				-0.03 (0.04)	-0.88	0.38
Topic [Holidays] - Leisure				0.02 (0.04)	0.60	0.55				-0.05 (0.04)	-1.33	0.19
Topic [Holidays] - Lifestyle				-0.06 (0.04)	-1.37	0.17				-0.05 (0.04)	-1.21	0.23
Topic [Holidays] - Work and education				-0.06 (0.04)	-1.66	0.10				-0.01 (0.03)	-0.20	0.84
Item type [blanks] -choose				0.17 (0.05)	3.32	<0.01				0.09 (0.07)	1.23	0.22
Item type [blanks] - match				0.19 (0.05)	4.12	<0.001				0.06 (0.07)	0.87	0.38
Item type [blanks] - MCQ				0.14 (0.05)	2.88	<0.01				0 (0.07)	0.01	0.99
Item type [blanks] - names				0.21 (0.05)	4.34	<0.001				0.08 (0.07)	1.08	0.28
Item type [blanks] - SA				0.13 (0.04)	2.94	<0.01				0.24 (0.07)	3.46	<0.001
Track length (sec)				0 (0)	-0.08	0.93				0 (0)	0.86	0.39
Speaker gender [both] - Female				-0.05 (0.03)	-1.97	0.05				0 (0.03)	-0.16	0.87
Speaker gender [both] - Male				-0.06 (0.03)	-2.16	<0.05				-0.06 (0.02)	-2.67	<0.001
Pseudo - R-squared	0.345			0.594			0.347			0.639		

Spanish Reading

Table D.5 Results of discrimination modelling of Spanish reading assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	-0.81 (1.07)	-0.75	0.45	1.2 (0.94)	1.28	0.20	-12.18 (2.88)	-4.22	<0.001	-15.08 (2.75)	-5.48	<0.001
Mean GCSE	0.24 (0.23)	1.06	0.29	-0.09 (0.2)	-0.43	0.66	2.03 (0.47)	4.36	<0.001	2.52 (0.44)	5.67	<0.001
Year [2017] - 2018	-0.09 (0.03)	-2.70	<0.01	-0.05 (0.03)	-1.34	0.18	-0.26 (0.07)	-3.51	<0.001	-0.37 (0.07)	-5.02	<0.001
Board [AQA] - Pearson	-0.01 (0.02)	-0.36	0.72	-0.02 (0.02)	-1.01	0.31	0.19 (0.06)	3.23	<0.01	0.27 (0.06)	4.57	<0.001
Board [AQA] - WJEC	0.1 (0.05)	2.12	<0.05	0.11 (0.04)	2.60	<0.05	0.1 (0.04)	2.28	<0.05	0.11 (0.04)	2.49	<0.05
language [English] - target	-0.02 (0.03)	-0.68	0.50	0.05 (0.03)	2.10	<0.05	-0.01 (0.03)	-0.45	0.66	0 (0.03)	-0.08	0.93
S1 (vocab difficulty)				-0.01 (0.03)	-0.27	0.79				-0.01 (0.03)	-0.19	0.85
S3 (grammar difficulty)				0.01 (0.03)	0.34	0.73				0.03 (0.03)	1.00	0.32
S6 (w ork required)				-0.1 (0.03)	-3.79	<0.001				0.01 (0.03)	0.57	0.57
Word Count				0 (0)	-0.32	0.75				0 (0)	-2.08	<0.05
Words per sentence				0 (0)	0.25	0.81				0 (0)	1.27	0.21
Lexical variety				-0.28 (0.17)	-1.66	0.10				-0.06 (0.16)	-0.38	0.70
Lexical density				-0.01 (0.11)	-0.08	0.93				-0.22 (0.2)	-1.10	0.27
Lexical unfamiliarity				-0.02 (0.11)	-0.22	0.83				-0.06 (0.14)	-0.46	0.64
Pictures included				-0.06 (0.03)	-1.91	0.06				0.01 (0.04)	0.15	0.88
Topic [Extract] - Home and environment				-0.06 (0.04)	-1.55	0.12				-0.04 (0.03)	-1.25	0.21
Topic [Extract] - Home and environment				-0.03 (0.04)	-0.77	0.44				0.04 (0.04)	1.00	0.32
Topic [Extract] - Leisure				0.01 (0.04)	0.26	0.79				0.03 (0.03)	0.91	0.36
Topic [Extract] - Lifestyle				-0.05 (0.04)	-1.43	0.16				0.03 (0.03)	1.09	0.28
Topic [Extract] - Work and education				0.01 (0.04)	0.18	0.86				-0.02 (0.04)	-0.59	0.56
Item type [blanks] -choose				0.05 (0.07)	0.71	0.48				0.02 (0.07)	0.36	0.72
Item type [blanks] - match				0.03 (0.03)	0.96	0.34				-0.02 (0.06)	-0.39	0.70
Item type [blanks] - MCQ				-0.1 (0.03)	-3.04	<0.01				-0.03 (0.06)	-0.57	0.57
Item type [blanks] - names				0.01 (0.04)	0.14	0.89				-0.07 (0.07)	-0.92	0.36
Item type [blanks] - SA				-0.02 (0.03)	-0.46	0.65				0.04 (0.06)	0.79	0.43
Item type [blanks] - Translation				0.39 (0.07)	5.42	<0.001				0.32 (0.08)	3.87	<0.001
Pseudo - R-squared	0.086			0.512			0.144			0.396		

Spanish Listening

Table D.6 Results of discrimination modelling of Spanish listening assessments.

	Foundation						Higher					
	Basic			Full			Basic			Full		
	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value	Beta (SE)	Z-value	p-value
(Intercept)	-0.17 (0.39)	-0.43	0.66	-0.28 (0.41)	-0.68	0.50	0.77 (1.07)	0.73	0.47	-2.25 (1.14)	-1.98	<0.05
Mean GCSE	0.1 (0.08)	1.21	0.23	0.1 (0.07)	1.35	0.18	-0.08 (0.17)	-0.49	0.63	0.36 (0.18)	1.99	<0.05
Year [2017] - 2018	0.02 (0.02)	0.97	0.33	0.03 (0.02)	1.18	0.24	0.05 (0.02)	2.19	<0.05	0.03 (0.03)	1.16	0.25
Board [AQA] - Pearson	0.03 (0.03)	1.08	0.28	-0.02 (0.03)	-0.59	0.56	-0.07 (0.04)	-1.97	<0.05	0.07 (0.05)	1.26	0.21
Board [AQA] - WJEC	0.07 (0.03)	2.08	<0.05	0.05 (0.03)	1.65	0.10	-0.11 (0.03)	-3.62	<0.001	-0.04 (0.04)	-1.06	0.29
language [English] - target	-0.06 (0.03)	-1.83	0.07	0.06 (0.03)	1.62	0.11	0.03 (0.03)	1.10	0.27	0.07 (0.03)	2.17	<0.05
Speech Speed (sec)	-0.04 (0.03)	-1.23	0.22	0.06 (0.04)	1.37	0.17	0.03 (0.05)	0.62	0.54	-0.04 (0.07)	-0.60	0.55
Pause length (sec)	0 (0)	-0.02	0.98	0 (0)	-1.74	0.08	0.01 (0)	2.75	<0.01	0 (0)	0.17	0.86
Time between tracks (sec)	0 (0)	0.10	0.92	0 (0)	1.90	0.06	0 (0)	0.58	0.56	0 (0)	0.97	0.33
S1 (vocab difficulty)				-0.05 (0.03)	-1.95	0.05				0 (0.03)	-0.02	0.98
S3 (grammar difficulty)				0.02 (0.03)	0.68	0.50				0 (0.03)	0.05	0.96
S6 (work required)				-0.04 (0.02)	-1.96	0.05				0.01 (0.02)	0.51	0.61
Word count				0 (0)	-2.53	<0.05				0 (0)	1.91	0.06
Words per sentence				0 (0)	-0.48	0.63				0 (0)	0.63	0.53
Lexical variety				0.19 (0.17)	1.11	0.27				0.24 (0.2)	1.19	0.24
lexical density				0.04 (0.09)	0.41	0.68				-0.38 (0.16)	-2.43	<0.05
Lexical unfamiliarity				-0.02 (0.08)	-0.30	0.76				0.16 (0.12)	1.32	0.19
Pictures included				-0.06 (0.03)	-2.13	<0.05				0.1 (0.04)	2.54	<0.05
Topic [Holidays] - Home and environment				-0.05 (0.03)	-1.73	0.09				0.1 (0.04)	2.92	<0.001
Topic [Holidays] - Leisure				-0.02 (0.03)	-0.66	0.51				0.04 (0.04)	1.24	0.22
Topic [Holidays] - Lifestyle				-0.03 (0.03)	-1.19	0.24				0.08 (0.03)	2.52	<0.05
Topic [Holidays] - Work and education				-0.04 (0.03)	-1.42	0.16				0.09 (0.03)	2.69	<0.001
Item type [blanks] -choose				0.16 (0.05)	2.95	<0.01				0.27 (0.09)	3.09	<0.001
Item type [blanks] - match				0.03 (0.05)	0.66	0.51				0.17 (0.08)	2.28	<0.05
Item type [blanks] - MCQ				0.01 (0.05)	0.10	0.92				0.05 (0.07)	0.76	0.45
Item type [blanks] - names				0.07 (0.06)	1.16	0.25				0.13 (0.08)	1.60	0.11
Item type [blanks] - SA				0.1 (0.05)	2.06	<0.05				0.24 (0.07)	3.27	<0.001
Track length (sec)				0.01 (0)	2.54	<0.05				-0.01 (0)	-1.81	0.07
Speaker gender [both] - Female				-0.02 (0.03)	-0.84	0.40				0 (0.03)	0.05	0.96
Speaker gender [both] - Male				-0.03 (0.02)	-1.40	0.16				-0.03 (0.03)	-1.15	0.25
Pseudo - R-squared	0.091			0.530			0.158			0.466		

Appendix E – Mark distributions

Reading

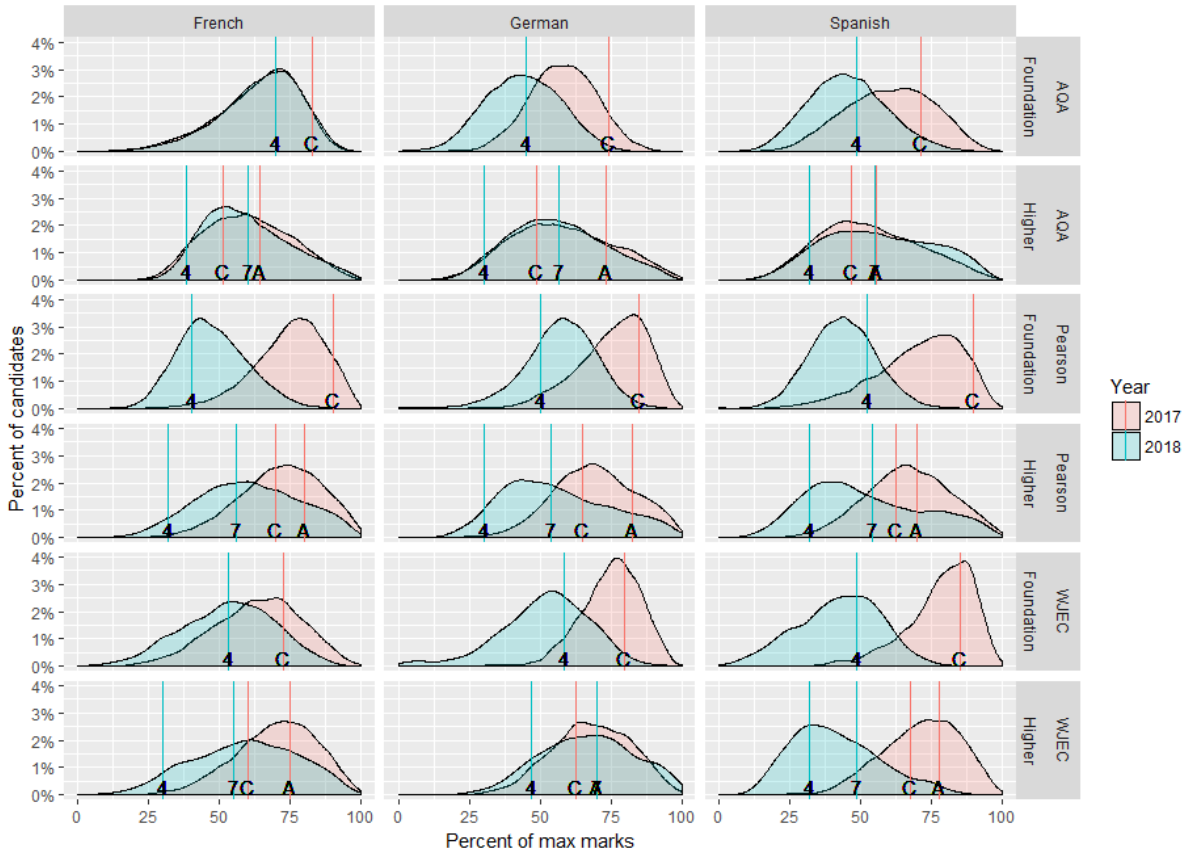

Figure E.1. Mark distributions for reading assessments in 2017 (red) and 2018 (blue).

Listening

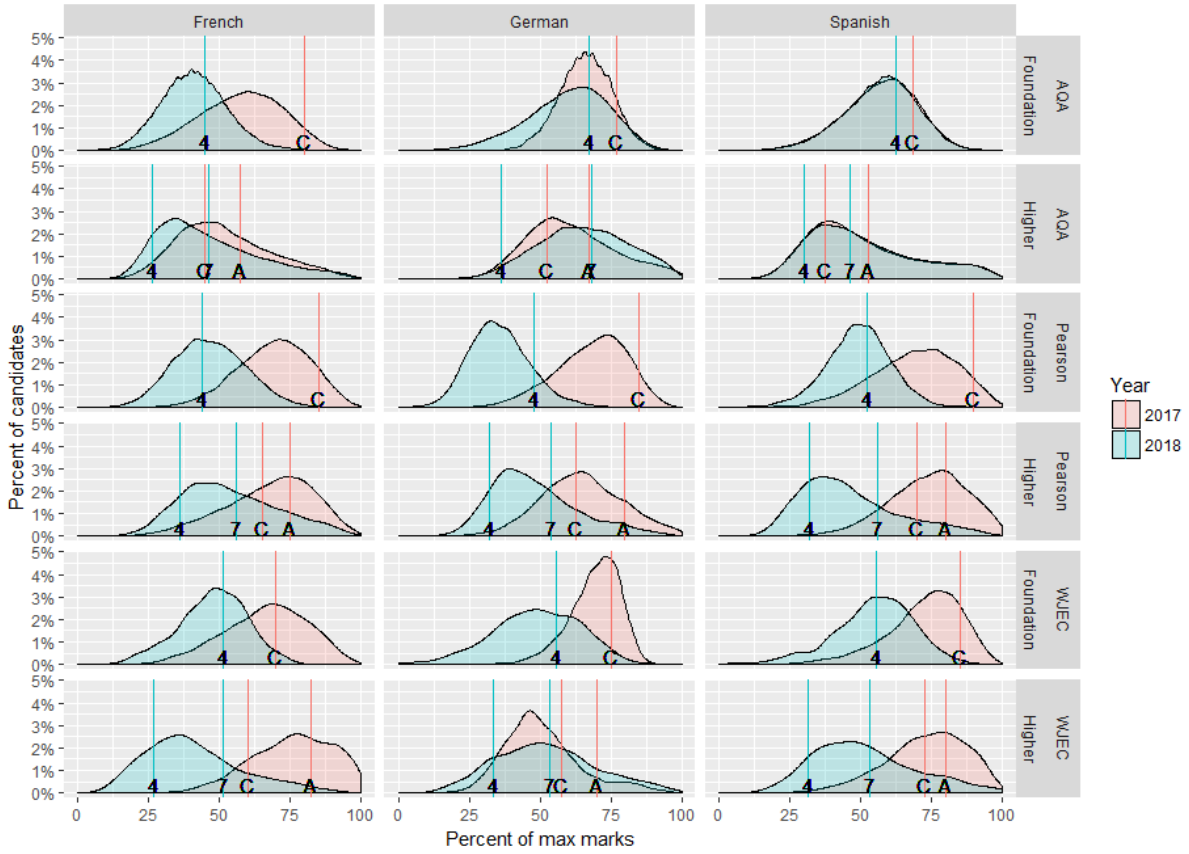

Figure E.2. Mark distributions for listening assessments in 2017 (red) and 2018 (blue).

Appendix F – Grade boundary changes

Table F.1. Change in grade boundaries from 2017 to 2018 as proportion of max mark. C/4 grade boundary.

		Listening		Reading	
Language	Board	Foundation	Higher	Foundation	Higher
French	AQA	-0.35	-0.19	-0.13	-0.13
	Pearson	-0.41	-0.29	-0.50	-0.38
	WJEC	-0.19	-0.33	-0.19	-0.30
German	AQA	-0.10	-0.17	-0.29	-0.19
	Pearson	-0.37	-0.31	-0.35	-0.35
	WJEC	-0.19	-0.24	-0.22	-0.16
Spanish	AQA	-0.06	-0.08	-0.23	-0.15
	Pearson	-0.38	-0.38	-0.38	-0.31
	WJEC	-0.29	-0.41	-0.37	-0.36

Appendix G – Component models

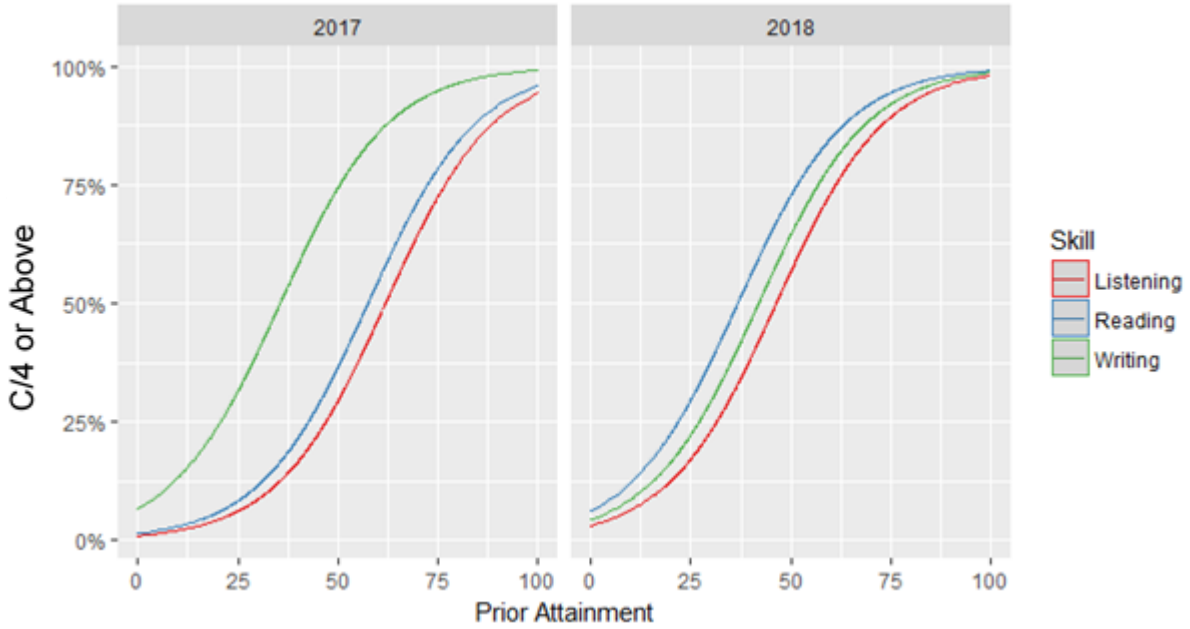

Figure G.1. Probability of attaining a C/4 (or above) in different components by prior attainment in German

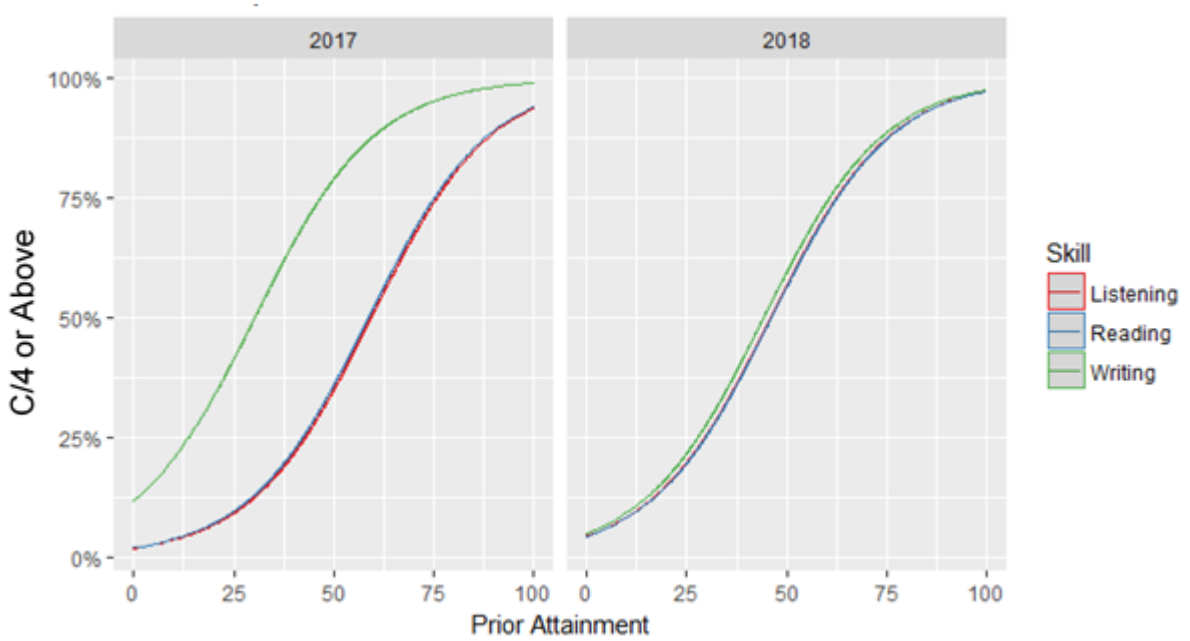

Figure G.2. Probability of attaining a C/4 (or above) in different components by prior attainment in Spanish

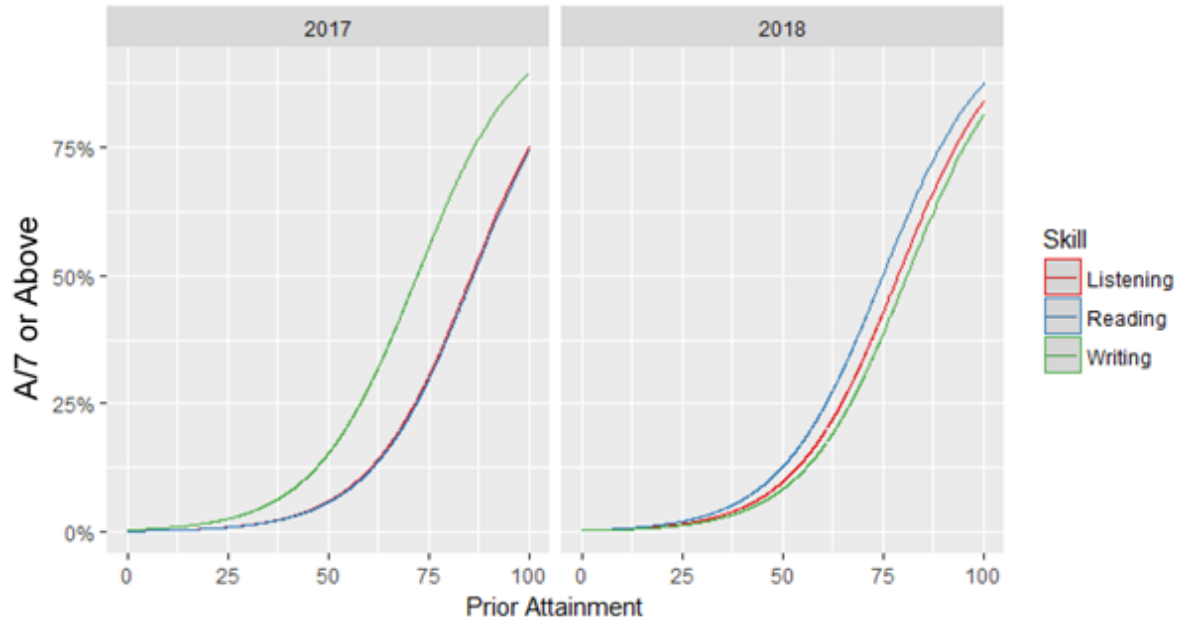

Figure G.3. Probability of attaining a A/7 (or above) in different components by prior attainment in German

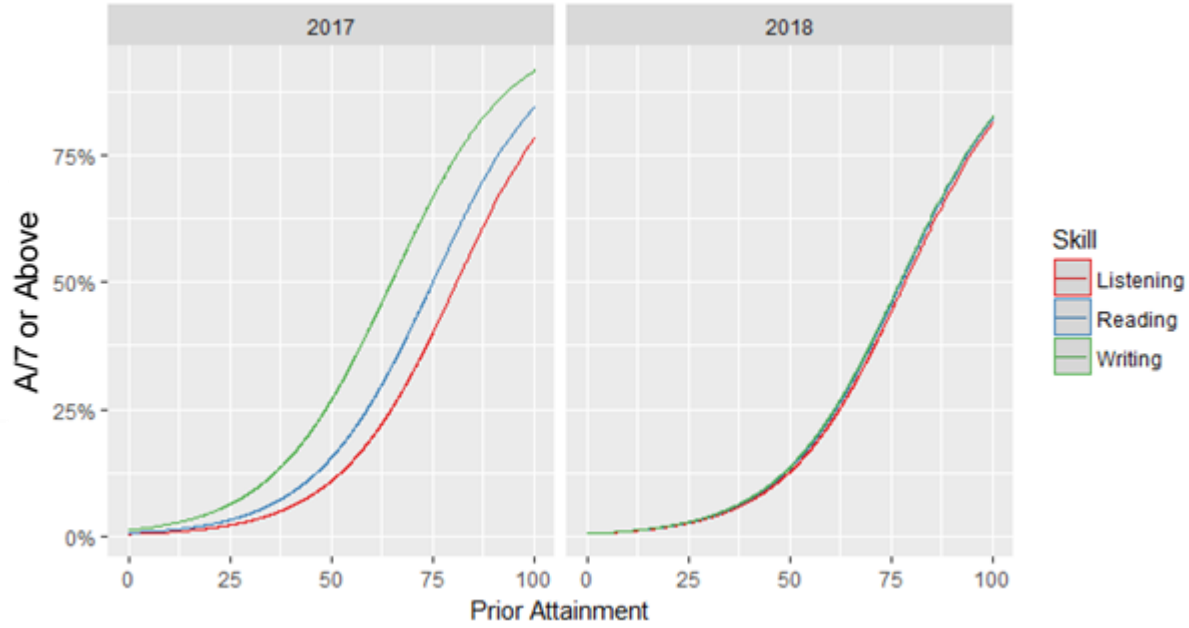

Figure G.4. Probability of attaining a A/7 (or above) in different components by prior attainment in Spanish

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

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

November 2019 **Ofqual/19/6557/2**