

**Mitigating the Effect of Language in the Assessment of
Science: A study of English-Language Learners in Primary
Classrooms in the UK**

Journal:	<i>Science Education</i>
Manuscript ID	SciEd-00391-2018.R2
Wiley - Manuscript type:	General Section
Keywords:	ELL, Assessment, Primary education, Science, Generalized linear model

SCHOLARONE™
Manuscripts

Mitigating the Effect of Language in the Assessment of Science: A study of English-Language Learners in Primary Classrooms in the UK

Abstract

Children coming from homes where English is not their first language constitute a significant and increasing proportion of classrooms worldwide. Providing these English-language learners (ELLs) with equitable assessment opportunities is a challenge. We analyze the performance of 485 students, both English-native-speakers (ENSs) and ELLs, across 5 schools within the UK in the 7-11-year age group on standardized summative Science assessment tasks. Logistic regression with random effects assesses the impact of English-language proficiency, and its interactions with question traits, on performance. Traits investigated were: question focus; need for active language production; presence/absence of visuals; and question difficulty. Results demonstrated that, while ELLs persistently performed more poorly, the gap to their ENS peers depended significantly upon assessment traits. ELLs were particularly disadvantaged when responses required active language production and/or when assessed on specific scientific vocabulary. Presence of visual prompts did not help ELL performance. There was no evidence of an interaction between topic difficulty and language ability suggesting lower ELL performance is not related to capacity to understand advanced topics. We propose assessment should permit flexibility in language choice and production type for ELLs with low English-language proficiency; while simultaneously recommend subject-specific teaching of scientific language begins at lower stages of schooling.

Keywords

ELL, assessment, primary education, Science, generalized linear model, random effects

1 | INTRODUCTION

The increase in migration has meant that a significant proportion of students in today's classrooms come from home environments where the dominant language is not English and who start, or continue, learning English on their entry to school. Within the UK, over the last fifteen years the number of these children, commonly known as English-language learners (ELLs), has nearly doubled in secondary schools (8% in 2000, 15.7% in 2015), and more than doubled in primary schools (8.7% in 2000, 20.1% in 2015) (Murphy & Unthiah, 2015; UK Department for Education, 2016). Similar trends have been seen in the US, where the number of ELLs in public/state schools has now reached an estimated 4.6 million students, making up 9.4% of the classroom (US Department of Education, 2017). In light of this growing demographic, and in line with the UN's convention on the right of all children to receive education that permits them to develop to the best of their abilities and talents (United Nations Committee, articles 28-30), addressing the particular educational needs of ELLs and providing them with equal opportunities to their English native speaking (ENS) peers is a significant and increasing challenge to educators worldwide.

Unfortunately, the assessment data to date, both national and international, shows that learners who are educated through the medium of a non-native language still tend to perform less well than their ENS peers (Honeycutt Swanson, Bianchini & Lee, 2014; Lyon, Bunch, & Shaw, 2012; Strand, Malmberg & Hall, 2015). The reasons behind this underperformance of ELLs across the curriculum are complex and likely to include linguistic, cultural and social effects, see Lee (2005) for a thorough review. However, there is considerable research that suggests that a significant factor may be 'the language difficulties' that students face in both learning and expressing their subject-specific knowledge and understandings (e.g. Author2 and Author1 2011; Rea-Dickins, Khamis, & Olivero, 2013). In Science specifically, amongst ELLs, achievement has been found to be more strongly associated with a student's language ability than with their gender, ethnicity, or economic status (Maerten-Rivera, Myers, Lee, & Penfield, 2010). Despite this achievement gap, there is evidence that Science is not inherently more difficult for such learners. Rather that their limited fluency restricts their capacity to produce clear scientific statements if required to do so in English (Curtis & Millar, 1988) and their abilities can hence be underestimated in assessment (Solano-Flores and Trumbull, 2003).

In this paper we consider the specific issue of providing valid and equitable summative scientific assessment of ELLs, identified by Lee (2005) as one of the most challenging problems in educational policy and practice. Many countries implement standardized tests at the end of primary and secondary school. The consequences of poor performance in these tests are highly significant for a learner, potentially influencing future opportunities and the direction of study at post-secondary education levels. Indeed, poor performance can affect a student's perception of themselves as a good Science, English or Mathematics learner. It may also lead to a student being streamed (separated by perceived ability) into a lower ability group or class, or indeed moved to a more vocational line of study that does not provide such an academically challenging curriculum (Evans, Schneider, Arnot, Fisher, Forbes, Hu, & Liu, 2016). Addressing this issue is therefore a key step if we wish to tackle the current under-representation of linguistically diverse learners in STEM post-secondary education.

Several suggestions have been made in the literature concerning the linguistic factors affecting the learning and assessment performance of ELLs, and what specific accommodations could be made to reduce the attainment gap between ELLs their ENS peers. In this article, we consider three facets of assessment that may impact on the extent to which learners are able to show how much they know in standardized assessment. These are: i) the use of visuals, ii) the requirement for students to actively produce language and iii) the specific focus of the assessment task. Much of the existing work in support of these facets has considered these aspects in isolation through single variable analyses. Conversely, our work seeks to jointly analyse their effects on the achievement gap between ELLs and ENSs whilst also controlling for alternative potential explanatory factors. Furthermore, some of the previous research into accommodations has concentrated more on their application in instructional settings whereas we specifically address their application in the context of assessment tasks. Finally, most previous work has focused on ELL achievement during secondary education, and is predominantly Europe, Australia, Sub-Saharan Africa or US based. Here, however, we locate this research at primary level in the UK where different factors may be of significance.

The identification of specific questions traits where ELLs underperform in summative assessment has significant and wide implications. Most directly, it may allow the provision of summative assessment tools which better separate scientific knowledge from language proficiency. However, we believe such information is also important to both formative assessment tools and pedagogy. In this context, it would enable teachers to identify particular aspects of Science where ELLs may need extra targeted support during the learning process. Furthermore, dependent upon the aims of the teacher, formative assessment may either be designed in an attempt to i) decouple language ability from understanding (potentially boosting scientific confidence in ELLs who may perform more strongly in such tests); ii) specifically assess improvements in the areas where ELLs are identified to struggle; or iii) teach strategies for these ELLs to employ, to enable them to better demonstrate their skills. This is especially important since some of the traits we consider are critical scientific skills which cannot be entirely removed from study of, and success in, the subject.

1.1 | Notation: education in the UK

In England and Wales, children between 7 and 11 are in Years 3 – 6, i.e. in the last four years of primary education. This period is also referred to as Key Stage 2. Throughout this time students work through a national curriculum that provides a uniform syllabus for schools covering all core subjects. At the end of Key Stage 2, at age 11, students' knowledge of this curriculum will be assessed via a statutory assessment known as SATs (Standard Attainment Tests) before they progress to the secondary phase. Table 1 gives an overview of the compulsory education system in the UK. The focus of our work is highlighted in grey.

Table 1: to be inserted here

2 | LITERATURE REVIEW AND RESEARCH QUESTIONS

We first provide an overview of the broad challenges facing ELLs during their science education and describe the difficulties involved in equitable assessment. We then review the literature in three specific areas identified as highly likely to impact upon the performance of ELLs, for which modified assessments have either been attempted previously or suitable modifications could be introduced. These are the focus for our study and drive our research questions.

2.1 | Science instruction and assessment for ELLs

When ELLs enter educational systems, they have to adjust rapidly to new academic, linguistic, cultural and social environments (Lee, 2004). Some of the skills these ELLs need to develop are no different from their ENS peers. For example, both groups of learners have to master subject-specific content and the specific academic language used to express this content. The nature of this required academic language proficiency is an area of considerable research with several competing conceptions and definitions, see Frantz, Bailey, Starr & Perea (2014), Flores & Rosa (2015) and Valdes (2004) for reviews. In its broadest sense, as described by Anstrom, DiCerbo, Butler, Katz, Millet & Rivera (2010, p. iv, cited in Schleppegrell, 2012, p. 409), it is “the language used in school to help students acquire and use knowledge”. In our research, we define the construct along the same lines in terms of the ability of a learner to draw upon their existing language resources (e.g. knowledge of syntax, lexis, academic registers, modality) to decode meanings from texts and to construct spontaneous responses to assessment questions to convey appropriate meanings.

On average it can take ELLs up to 7 years to develop their academic language proficiency and, even after ELLs are reclassified as English proficient, they may still need help to refine their academic language skills (Siegel, 2007). To effectively develop academic language proficiency, it is argued that ELLs need explicit, intensive and ongoing support from their teachers (Hammond 2014; Kieffer, Lesaux, Rivera, & Francis, 2009). Receiving this support is important as highly developed academic language proficiency is a crucial factor in determining success in high-stakes end of school examinations (Murphy & Unthiah, 2015).

In Science, academic language proficiency is seen as forming a part of scientific literacy. Scientific literacy is a complex construct. It entails knowledge about the field (i.e. subject-specific content), genre (the global patterns of text organization that package this knowledge), and unique scientific lexicon and semantics (Martin, 1993; Lemke, 1990; Fang, 2006). For learners to become scientifically literate, all these components need to be explicitly taught to them. However, until recently this has not been common practice as subject-teachers often lack the training and expertise to teach language and scientific literacy skills as part of their subject-specific lessons (Martin, 1993; McCloskey, 2002). Where subject and literacy integrated teacher training has been introduced, positive effects on ELLs’ performance have been reported (Bravo, Mosqueda, Solis, & Stoddard, 2014; Shanahan & Shea, 2012; Lara-Alecio, Tong, Irby, Guerrero, Huerta, & Fan, 2012). The language of Science, however, does not only pose difficulties to ELLs but also, and often equally, to learners who speak English as their first language (Fang, 2006; Wellington & Osborne, 2001). Moreover, it is not only specialized scientific lexis that can pose comprehension challenges, but also ordinary words - such as ‘school’, ‘volume’, ‘power’, ‘heat’ - when used in metaphorical ways where they have additional meanings of which learners are unaware (Fang, 2006; Gee, 2008; Fung & Yip, 2014). Both scientific and everyday words carrying subject specialized meaning also become a

challenge to learners when used in assessment tasks, especially when learners are assessed through the medium of a language that is not their mother tongue.

There is a long-standing debate on the use of language in assessment. Ideally, any subject-specific assessment should aim to distinguish, as far as possible, subject-specific knowledge from English-language proficiency. Abedi (2004) argues that ELLs' performance may be underestimated due to confounding of language and content, as any test that employs language, in part, also measures language skills (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999). Researchers have thus identified a need to develop testing that allows a more valid assessment of ELLs' subject-knowledge (Kopriva, 2008; Pitoniak, Young, Martiniello, King, Buteux, & Ginsburgh, 2009; Solorzano, 2008). One proposed approach to mitigate for language is to assess ELLs using tests items given in both English and their home language(s) (Solano-Flores & Trumbull, 2003; Buxton et al, 2014). However, this approach is not without opposition, see Abedi (2002), who argues that, as well as adding considerable expense, such translation may introduce ambiguities and generate exams for different cultures and languages that are not truly equivalent. Furthermore, in many English-medium educational contexts, use of other languages for learning and assessment is either not allowed (Proposition 227, 1998) or not encouraged (Lee, 2005). This is despite extensive evidence showing that restricting use of ELLs' native language can hamper learners' understanding of important concepts (Bunyi, 1999; Cleghorn, Merritt, & Abagi, 1989), while its use can help develop learners' subject-specific knowledge (Garcia, 1997; Ryu, 2015). Awareness amongst teachers of the potential benefits of using ELLs' native language however remains low (Honeycutt Swanson, Bianchini & Lee, 2014).

2.2 | Specific factors affecting learning and assessment for ELLs

Several specific factors have been identified in the literature as having a particular impact upon the scientific education of ELLs, we consider three in particular: the use of visuals; requirement for the student to actively produce language; and assessment focus. The extent to which these factors have already been studied in the assessment setting varies. Some have already been investigated as specific assessment accommodations, although often analysed without comparison to equivalent ENS students or as single variables without controlling for other covariates and confounding effects. Others have mainly only been studied in the learning setting but could be integrated into assessment design.

Use of Visuals

Several authors have proposed that systematic use of models and visuals in Science lessons may help ELLs by providing concrete representations of abstract ideas and complex relationships (Department for Education and Skills: 2002; Buck Bracey, 2017). Visual tools may also help learners to decode or visualize the language of scientific texts, or - where the language is the point of departure - to gloss the images that complement the text (Unsworth & Cleirigh, 2009).

Much previous research in this area concentrates on improving understanding during the learning process rather than specifically on assessment provision. In instructional settings, use of visualizations has been shown to improve understanding in Biology (Kiboss, Ndirangu & Wekesa, 2004), Chemistry (Ardac & Akaygun, 2004) and Science (Cromley, Weisberg, Dai, Newcombe, Schunn, Massey & Merlino, 2016). Furthermore, this previous work often studies the effect of introducing visuals to either ELLs or ENSs separately (e.g. Lara-Alecio, Tong, Irby, Guerrero, Huerta, & Fan, 2012), or does not distinguish between them (e.g. Ardac & Akaygun, 2004; Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel & Metcalfe, 2007); as

opposed to whether their introduction may have a differential effect on the two groups (ELLs and ENSs) that affects the size of the performance gap. These studies generally suggest that visuals offer most benefit when used alongside interaction with a more capable other (teacher or another learner) (Yip, 2004); or alongside learners' active production of scientific language (Barnett, 1992).

Investigations into the effect of incorporating visuals into assessment are somewhat less developed. In particular the strength of evidence for an interaction with language is unclear. In light of the work within a learning context, visuals may only help in assessment if students are able to make *independent* use of them alongside the text – this may not be possible for learners with low language proficiency. Siegel (2007) undertook a study modifying written test items, including adding visual supports, on middle school students (K8). She identified significant improvements from the pre- to post-modification items for both advanced ELLs and ENSs separately but found no statistically significant evidence for an interaction between modification and language, suggesting no evidence that these modifications were particularly useful for the ELLs. She concluded however that this may have been down to the small sample size. Solano-Flores (2014) and Wang (2012) have recently proposed, from a semiotics perspective, that use of visual illustrations in assessment protocols may make a difference to the relative performance of ELLs as long as they have certain skills (unrelated to content knowledge) to use them. They also suggest that, due to cultural differences, different communities may require different illustrations to aid them. However, the level of empirical support for this is currently unclear. Wang (2012) focused mainly on patterns of correlations or multiple independent ANOVAs for which the determination of statistical significance is hindered by the issue of multiple testing; while Solano-Flores et al. (2014) found, in their pilot study, no evidence for an interaction between language and presence of illustrations in determining assessment performance. These investigations have also focused on middle school education rather than primary.

In our study, we are not able to consider all potential variables one might manipulate in the use of illustrations in assessment tasks. As a first step towards addressing whether, after accounting for other potential explanatory factors, illustrations can help reduce the performance gap in assessment, we examine the effects of the presence or absence of visuals. Possible areas for further study regarding visuals are discussed in our limitations section.

Need for active language production

Another factor that may affect learners' ability to demonstrate subject-specific knowledge fully in assessment settings is the need to produce language actively as part of their response (Brown & Spang, 2007; Duran, Dugan & Weffer, 1998). It is this creation of spontaneous responses and appropriate meaning making in learning tasks and assessment questions that, for the purposes of our work, we call 'active language production'. More detail on the types of tasks we consider as requiring active language production can be found in Section 3.4 and examples of these tasks can be found in Supplementary Material 2.

In the context of instructional settings Rainey, Maher, Coupland, Franchi and Moje (2017), drawing on Moje's (2015) 4-Es heuristic model for disciplinary literacy teaching, provide very helpful examples of teaching practices that aim to facilitate the development of learners' disciplinary knowledge and literacy through active engagement with, and analysis of, its content and discourse. The 4-Es in the model are described as: "1) engaging students in work that aligns with the problem- and text-based work of disciplinarians, 2) eliciting and engineering students' learning opportunities so they are able to successfully accomplish classroom tasks and learn disciplinary practice from them, 3) examining words, language, and representations, and 4) evaluating words and ways with words within and across domains" (Rainey, Maher, Coupland, Franchi & Moje, 2017: p. 372).

Furthermore, a theoretical framework on science literacy and language use by Wallace (2004) unpacks the construct of active academic language production through “Authenticity”, “Multiple Discourse” and “Third space” dimensions, all of which are presented on a continuum. The Authenticity dimension represents ‘expression’ and argues for “gradual incorporation of more scientific vocabulary, syntax, functional grammatical elements [...] into a student’s [originally vernacular/everyday] written and verbal expression” (ibid: 911-912). The Multiple discourse dimension represents ‘voice’ and signifies progression from private genres of speculative discourse to public genres of evidence-based scientific discourse. Finally, the Third space dimension represents ‘meaning’ and ‘signifies the personal and individual construction of language between two participants in a discourse’ (ibid). It is this last dimension that is particularly important in enhancing learners’ *active* use of language in the classroom as it allows for active probing of, and experimentation with, wide range of verbal and written discourses by the students and teacher while unpacking mutual meanings. Arguably, learners’ successful engagement with the elements of Rainey, Maher, Coupland, Franchi and Moje’s (2017) model and the dimensions in Wallace’s (2004) framework during instructional settings may also enhance their ability to articulate knowledge and understanding more successfully in assessment settings.

Most previous investigations in the area of active language production concentrates on the learning environment but, if it is considered significant, suitable accommodations could potentially also be introduced into *early* educational assessment, as a way to better prepare learners for the demands of formal assessment tasks in later stages of education. In instructional settings, Robinson (2005) emphasizes the importance of providing ELLs with opportunities to produce language actively, to negotiate the meaning of scientific terms and to construct their own understandings of the words. The author argues that without this, learners’ understanding of key subject-specific vocabulary, and hence their ability to talk about the subject, may remain underdeveloped. Honeycutt Swanson, Bianchini and Lee (2014) however report that fluent ENSs are nearly three times more likely to participate in whole class conversations than ELLs. Robinson (2005) also suggests that such low participation may be due to limited English vocabulary preventing ELLs from producing active oral contributions.

Assessment tasks in Science, especially those in later stages of schooling, routinely require active production (creation) of language. Having limited proficiency in this area may penalize learners, particularly ELLs. To further investigate this assertion, we therefore include in our study an analysis of the relative effect of assessment tasks which require active (i.e. spontaneous) language creation in their solution, in comparison to those requiring only passive reproduction of language (i.e. incorporation or transferring of provided linguistic models into responses), on the performance gap between ELLs and ENSs.

Focus of Assessment Task

Finally, different formats or wording of assessment prompts have also been seen to affect ELLs’ performance (Routitsky & Turner, 2003; Abedi, 2002). Shaw (1997) discusses how different foci may explain variations in ELLs achievement, suggesting performance on tasks requiring dependence on text is more significantly affected by language than those requiring graphs or calculation of formulae. Furthermore, Dempster and Reddy (2007) investigate how sentence complexity, specifically the use of unfamiliar and long words differentially affects ELL and ENS learners’ performance on multiple-choice questions in Mathematics and Science leading to poorer performance for learners who had limited English-language proficiency. However, whilst there is considerable research in the language teaching, pedagogy and assessment fields with regard with the impact of task type on learner performance, we were unable to locate research focusing on the specific types of knowledge (scientific and linguistic) that we believe comprise most forms of Science assessment tasks at

primary school level between the ages of 7-11. These are: understanding of research procedure (R); understanding of scientific fact (SF); production or recognition of scientific vocabulary (V); and understanding of scientific fact in combination with the production or recognition of scientific vocabulary (SFV).

Finally, we note that the nature of science assessment items is continuously evolving, in many cases away from recall and towards other methods thought to better demonstrate understanding e.g. the Next Generation Science Standards (NGSS) in the United States. With this conceptual shift is an accompanying linguistic shift. National Research Council (NRC, 2012) categorises three dimensions of the NGSS framework from ‘scientific practices’ to ‘crosscutting concepts’ to ‘disciplinary core ideas’. We consider our ‘research procedure’ (R) question type to be similar to ‘scientific practices’ and ‘understanding of scientific fact’ (SF) - to be similar to ‘disciplinary core ideas’, as specified in the NGSS framework. Lee, Quinn & Valdes (2013) exemplify further relationships and convergences between the disciplines of Mathematics and Science and the Standards for English language arts.

2.3 | Research Questions

We seek to quantitatively assess the joint effect of these three factors through a designed experiment while also controlling for potential external factors. Joint analysis is important since several of these factors may typically intersect in test items. Our focus is not primarily on whether these factors of assessment design make an overall difference to performance but rather if they have differential effects for ELLs and ENSs (i.e. interact with language) and so alter the observed gap between the two groups. Specifically, we aim to examine the following research questions:

1. Does ELL status impact upon achievement in primary school assessment of Science? If so, how? What is the performance gap between ELL and ENS students?
2. Does altering the traits/styles of assessment question (specifically through the use of visuals, altering the requirement for active language production, and choice of question focus) differentially affect ELLs’ and ENSs’ performance? Which of these aspects increase the performance gap between ELLs and ENSs; and which reduce it?

A suite of assessment questions designed to systematically vary across the proposed three factors is used to infer, via logistic regression with random effects, their relative effect on the performance gap between ELLs’ and their native speaking peers. We control for additional factors of topic difficulty, school ELL density and student age which may also affect performance; as well as the covariance in responses introduced through the multiple responses from the same students and schooling. See Section 3.4 for details on study design.

3 | DATA AND METHODS

3.1 | Geographical context

In the UK, increases in the ELL population are not evenly dispersed. Our study was conducted in schools of the Yorkshire and Humber region. Yorkshire and Humber is one of the most heavily ELL populated regions in the UK with 157 schools having more than 50% of their learners being ELL (Strand, Malmberg & Hall, 2015: 5). This lies behind only London (919 schools with more than 50% ELLs) and the West Midlands (201 schools). Furthermore, Yorkshire and Humber ranks poorly in national rankings of student attainment, coming second

lowest of the ten English regions. When ranked in terms of the statutory General Certificate in Secondary Education school examinations at age 16, only 63.8% of pupils in Yorkshire and Humber achieved the desired five grades at the top (A* to C) levels (Department for Education, 2016).

3.2 | Study background

This paper reports on a study conducted over a 2-year period (September 2013-August 2015) in five state primary schools in an inner city area. These schools were selected on the recommendation of a senior ELL consultant from the Local Authority. The research had two phases: a pre-intervention baseline study, and teacher and materials intervention study. In this paper we present only the findings from the baseline study.

The schools had varying densities of ELLs, ranging from 17% to 96%, and represented children from various ethnic, social and economic backgrounds (Supplementary Material 1, Table S1). In each school, one class from each year group in Key Stage 2 (ages 7-11) was selected, totaling four classes per school, and eighteen classes for the entire project. Classes were selected by the schools' headteachers based on the teachers' willingness to participate.

3.3 | Participants

A total of 485 primary school children, 120 parents and 29 teaching staff took part in the baseline study. Only learner data is reported in this paper. Table 2 provides a breakdown of learners by school and year group.

Table 2: to be inserted here

3.4 | Research framework and question design

Our assessment framework is shown in Figures 1 and 2. Figure 1, provides an overview of the topics for each year/age group.

Figure 1: to be inserted here

Several topics were selected for each year group, with two topics for Year 3 (7-8 years old); four topics for Year 4 (8-9 years old); and five topics for Years 5 (9-10 years old) and 6 (10-11 years old). In addition, there was an overlap for some topics between the year groups. This meant that each year group, apart from Year 3, was assessed on at least one topic from the preceding year, as well as their year-specific topics. Additionally, all four year-groups were assessed on the topic of "Growing plants" to facilitate future analysis (beyond the scope of this paper). Four questions were set for each topic.

Figure 2 specifies the individual question characteristics which were designed to vary according to the three specific factors identified in the theoretical review. These were 1) focus, 2) requirement for active language production, and 3) presence of a visual aspect. With our eight topics, this led to thirty-two assessment questions in total, with each learner, depending on Year group, completing between eight and twenty questions. All assessment questions were taken from the 2003 – 2011 National Curriculum assessment past papers (Qualifications and Curriculum Authority, 2003-2011) and are, thus, representative of the actual examination. Since topics varied in difficulty this was also included as a covariate for analysis.

Figure 2: to be inserted here

Figure 2 shows that each question had the following series of traits potentially affecting student performance:

- *Focus* – defined by question with four groups: understanding of research procedures (R) (see Supplementary Material 2, Image S1); understanding of a scientific fact (SF) (Image S2); production or recognition of scientific vocabulary (V) (Image S3); understanding of a scientific fact *and* production or recognition of scientific vocabulary (SFV) (Image S4). The questions focusing on research procedure (R) were fewer in the overall assessment corpus, but we decided nonetheless to include them in our analysis as they incorporated the core traits that we were aiming to investigate.
- *Visual* – defined by question: either the question used visuals (Images S1 and S2) or did not (Images S3 and S4).
- *Language Production* – defined by question: either the solution to a question required *active production* of language or *no production (passive reproduction)* of language.
- *Difficulty* – defined by topic with three levels in line with the National Curriculum for Science at Key Stage 2: least conceptually demanding, taught at the lower stages of the primary Science curriculum (one beaker); more demanding, taught at the middle stages of the primary Science curriculum (two beakers); and most demanding, taught at the higher stages of the primary Science curriculum (three beakers).

Active language questions covered various types of tasks: (N) name a process/fact (Image S3); (E) explain a process/phenomenon (Images S2 & S4); (N&E) name and explain (Image S5); and (DB) describe a process using personal linguistic resources (Image S6). Passive reproduction tasks included: (D) demonstrate understanding via drawing (Image S1); (T/F) decide if a statement is true/false (Image S7); (L) label a diagram using labels provided (Image S8); (CD) complete a diagram using information explicitly provided (Image S9); (T) tick the correct response from those provided (Image S10); (M) match the facts provided in a specific way (Image S11); (Y/N) respond Yes/No according to whether facts provided are accurate (Image S12).

3.5 | Data collection

The assessment tasks were undertaken at the beginning of the 2013-4 academic year. To help identify aspects that learners found particularly problematic, learners were invited to circle or underline any words that they found unfamiliar or problematic. We also invited non-native English speaking learners to complete - optionally - their assessment papers, or individual tasks, in their first language/s. The instructions given to the learners are presented in Supplementary Material 3, Figure S1. Finally, because standardized data on learners' language proficiency was unavailable to us through the schools, we invited learners to complete a language background questionnaire (see Supplementary Material 3, Figure S1). These data enabled us to make informed decisions on their English language proficiency classification (see 3.6).

3.6 | Data analysis

We assess whether the capacity of a student to answer a question correctly was dependent upon language ability and question trait (i.e. focus, requirement for active language production,

presence of visual aspect, difficulty); and significantly potential interactions between these factors. Analysis was performed via a generalised linear mixed model (Bolker et al., 2009).

Classifying English language proficiency

Following the analysis of learner language background questionnaire data, the learners' language proficiency was categorised into three classes of descending ability:

- 1) 'Native English' – self-reported English as their first language AND named no other language/s as being spoken at home,
- 2) 'ELLLevel1' – self-reported English as their second or third language by either stating this explicitly and/or by naming one or more 'other' languages as those spoken at home. These learners also self-reported speaking English 'very well',
- 3) 'ELLLevel2' – as 'ELLLevel1' but either self-reported speaking English 'OK' or 'Not very well'.

Learners who reported speaking English as their second or third language but did not specify their proficiency level were removed from analysis – 22 cases. Learners with undecipherable language responses were also excluded. Table 2 shows the split of learners by language class.

Correctness

Several questions had multiple parts enabling a student to be either entirely correct, partly correct, or entirely incorrect, in accordance with the scoring specifications detailed in official SATs marking scheme guides for KS2 Science (Supplementary Material 4). Below we present the analysis considering "entirely correct" as the outcome of interest. However, we also performed a separate analysis considering "either partly or entirely correct" as the response variable. Little difference was seen between conclusions (Supplementary Material 5).

3.7 | Initial data summary

We observed 6680 individual question responses. Table 3 shows the number of students by language proficiency class assessed on each topic.

Table 3: to be inserted here

Figure 3 presents maximum likelihood estimates (solid lines in middle of boxes) together with 95% confidence intervals for the proportion of students in each language proficiency class answering each question entirely correctly. Topics 1 and 2 (teeth and eating; and growing plants) were generally answered well except for question 2c which was answered considerably more poorly, but consistently, across the cohort. Learners struggled on the later parts of topic 3 (magnets) and all of topic 4 (particularly 4d) on habitats and the food chain. Learners generally did well on topics 5 and 6 (solids and liquids; and changing states) but more poorly on the high difficulty topics 7 (sounds) and 8 (circuits).

Figure 3: to be inserted here

Language proficiency also appears to affect performance. For most questions, the thicker solid lines indicating the proportion of students in the study who answered correctly, lie higher for Native English speakers (red) than the ELLs. While this is most evident for those students with

lower levels of English proficiency (ELL Level 2; blue), even amongst ELLs who speak English “very well” (ELL Level 1; green), a lower proportion answered correctly compared to native English speakers for the large majority of questions. Significantly however, the amount by which native English speakers outperform their ELL peers appears to vary considerably dependent upon the question. ELL performance on questions 2b, 2d and 5a-d are greatly worse than the Native English speakers. Conversely questions 1d, 2a, 3b and 4b-c show a much reduced performance gap, if at all. This supports our hypothesis that there may be certain traits of assessment questions that cause ELLs to particularly struggle, while for others they are at less of a disadvantage, as we discuss in Section 4.

3.8 | Analysis approach

Logistic regression with random effects

To investigate how the ability to answer an assessment question depends upon language ability and question traits, while accounting for the various year groups and schools, we model the probability p_{ij} of student i answering question j correctly (or partly correctly) via a logistic regression with random effects:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \boldsymbol{\beta} \cdot \mathbf{X}_{ij} + Z_k,$$

where \mathbf{X}_{ij} denotes the fixed properties of student i and question j and $Z_k \sim N(0, \sigma_k^2)$ are random effects describing shared influences between responses. This enables us to investigate the combined effect of our potential explanatory factors: 1) Fixed effects relating to school/student: student year, school ELL density, four level English proficiency; 2) Fixed Effects relating to question: topic difficulty, focus, presence of visual aspect, need for active language production; 3) Random effects: student, school, question (i.e. 1a, 1b, ..., 2a, ...).

Random effects are needed since, for example, each student provides multiple responses that are expected to be dependent – an able student would be more likely to get all of their questions correct than a less able student. Inclusion of a random effect intuitively provides some measure of individual student *ability* (separate from language, age, ...) with each student's *ability* considered to be drawn randomly from the entire population. Through random effects we can incorporate such “within student” correlation and make inference beyond our specific sample to the wider population. Similarly, random effects for school and question account for dependence in responses within a school (perhaps due to teacher/catchment) or that different questions on the same subject material might actually be of varying difficulty. Our random effects were treated as independent.

Model fitting and selection

An initial model was fitted with fixed effects of learner year group (Year); school ELL density (Density); and interactions between English proficiency (the three level ProfClass described earlier) and the four question variables – Topic difficulty (QuDifficulty), need for active language production (QuActive), Focus (QuFocus), and presence of a visual aspect (QuVisual). Initial random effects were school, student and question. Model selection was performed using AIC where both fixed and random effect terms were considered as possible terms to be dropped. After comparison of various possible models, our final model was:

$$\text{Correct} \sim \text{Year} + \text{Density} + \text{ProfClass} + \text{QuDifficulty} + \text{QuActive} + \text{QuFocus} + \text{ProfClass} * \text{QuActive} + \text{ProfClass} * \text{QuFocus}$$

$$+ (1 | \text{StudID}) + (1 | \text{QuestionID})$$

Full model output can be seen in Supplementary Material 5 together with model checking. Analysis was performed in R (R Core Team, 2017) using the lme4 (Bates, Maechler, Bolker, & Walker, 2015) library. Since we only had three random effects, a Laplace approximation was used for parameter estimation.

Significance of variables in final model

The large number of levels per random effect (440 students and 24 questions), together with the large overall number of responses compared to the number of fixed effect levels, suggest reliable p-values can be obtained via a likelihood ratio test. These are shown in Table 4 and provide very strong evidence that all these terms are significant in affecting the ability of a student to answer a question correctly.

Table 4: to be inserted here

How to interpret regression model results

Logistic regression models provide estimates of log-odds ratios for the different explanatory variables (see e.g. McCullagh and Nelder, 1989 for an explanation). In a scenario described by explanatory variables x , the odds of answering a question correctly are defined as:

$$\text{Odds}(\text{Correct} | x) = \frac{\text{Probability}(\text{Answer Correctly} | x)}{1 - \text{Probability}(\text{Answer Correctly} | x)}$$

Given two sets of scenarios you wish to compare described by explanatory variables x_1 and x_2 , the log-odds ratio from x_1 to x_2 is

$$\log(OR) = \log \frac{\text{Odds}(x_2)}{\text{Odds}(x_1)}$$

This allows you to compare how changing the nature of a question (or school, student...) will affect the probability of answering correctly. If the log-odds ratio is positive a student is more likely to respond correctly in scenario x_2 than x_1 . If it is negative, a student is less likely to respond correctly in scenario x_2 than x_1 .

A summary of the full model output with log-odds ratios can be found in Supplementary Material 5. However, the interaction terms in our model means direct interpretation of these values is somewhat difficult. We therefore describe our main findings by presenting, in tables 5 and 6, the log-odds ratios (and std. error) for all the significant language and question trait interactions compared to a suitable baseline. These tables should be interpreted in two ways. Firstly, the absolute values show how each interaction group performs compared to the baseline. Secondly, tracking down each column individually, the spread of values across the different language proficiencies (i.e. from native speaker down to ELL level 2) for each fixed question factor indicate which question traits reduce/increase the performance gap due to language. Those question traits with a narrower spread between English native speakers (ENSs) and ELLs have smaller differences in performance across the various language proficiency classes (assessed in terms of the odds). Finally, we also discuss those traits that we might have expected to affect responses but did not appear to play a significant role i.e. were discarded in our model selection.

4 | FINDINGS

In this section, we interpret our findings to address the research questions identified in Section 2.

4.1 | Log-odds ratios for interactions between question traits and language proficiency

Table 5 illustrates the effect of varying the question focus for students with different levels of language proficiency. We provide the log-odds ratio of answering correctly for each combination of question focus and language proficiency when compared to a baseline of a native English speaker assessed on knowledge of “scientific fact”. A positive number means a question is more likely to get answered correctly with that combination of variables (i.e. language proficiency and question focus) compared to the baseline; a negative number less likely. Here we assume the question requires active language production.

Table 5: to be inserted here

Table 6 demonstrates the effect, on students with differing language proficiencies, of varying the question style from one which requires active language production to one which does not. Here the question focus is fixed to be a “scientific fact”.

Table 6: to be inserted here

4.2 Research Question 1: Do ELL and English native-speaking learners perform differently in primary school Science assessment?

Looking down the columns in Tables 4 and 5, we see language proficiency has a large effect on performance. ELLs perform less well than native English learners whatever the question trait. For all the four question foci, and for both active and passive language categories, the estimated log-odds ratios for the ELLs lie below those for the corresponding native English learners. Furthermore, the consistent decrease in log-odds down the language proficiency classes within each column shows that the lower the level of English proficiency the larger the gap to the native English learners. The difference in log-odds between Native English learners and ELL level 1 is sometimes small suggesting these ELLs who can speak English “very well” may not be too disadvantaged. However, the larger differences between native English learners and those in ELL level 2 suggest that those who are only able to speak English “OK” or “Not very well” are likely to be affected to a much greater extent.

The size of the performance gap between the language proficiencies does however vary significantly dependent upon question focus and requirement to create active language. The performance gap between ELLs and ENSs is considerably increased if active language is required to answer a question compared to when it is not. Similarly, certain foci widen the differences between the ELLs and ENSs while others narrow it. We discuss this, as well as the influence of the other question traits on relative performance, in answer to our second research question.

4.3 | Research Question 2: What assessment traits interact with language ability to determine performance

All learners (both ELL and ENS) performed better when questions did not require *active language production* as (part of) the answer. For all language proficiency classes, the log-odds ratios on the right hand side of Table 6 (no active language production) are higher than their corresponding estimate on the left hand side (with active language production). However, the

requirement for active language does not affect the groups evenly. When active language is not required in a task response, the relative gap between ELL and ENS learners is much smaller than when active language is required. For a question assessing knowledge of scientific fact, the log-odds ratio (performance gap) between ENS learners and ELLs who spoke English at best “OK” (i.e. level 2) is -1.05, with a 95% CI of (-1.58, -0.53) if the question requires active language; but only -0.32 (-0.88, 0.23) if it permits passive language reproduction. ELLs (in particular those less proficient at English) are therefore able to perform more closely to their ENS peers when questions do not require active language production. Removing an active language production component from tasks may therefore reduce the achievement gap. However, even without the requirement for active language production, ELLs will still be expected to perform more poorly than their ENS peers.

Table 5 also shows that students generally performed best on the questions focusing on scientific research (R). ELLs who spoke English ‘very well’ (ELL level 1) do not perform much differently than ENS learners when the question focusses on scientific fact (SF); fact and vocabulary (SFV); or research (R). This suggests that, if questions are phrased suitably, ELLs with high level English skills can demonstrate similar levels of subject-specific knowledge to their ENS peers. ELL level 2 learners show larger performance gaps compared to the ENS group for all foci. However, when assessment tasks target *scientific* vocabulary (V) only, we see a clear difference between the relative performance of the language groups with the achievement gap between ELLs (both levels) and their ENS peers significantly increasing. While ENS learners perform better on a question focused on vocabulary than a question targeting scientific fact, both ELL learning groups perform worse. Moreover, the poorer the level of English, the larger the penalty compared with native English speakers. A lack of subject-specific vocabulary can thus heavily penalize ELLs if they are required to use such language in assessment even if, as indicated by the smaller gaps for other question foci, they have the underlying knowledge. We explicitly demonstrate this using some individual student responses in Section 6.

It is worth noting that the majority of Science assessment at Key Stage 2 does not require scientific/academic language. As long as questions are answered conceptually correctly, marks will be awarded. Hence, at the primary stage of education, English native speakers and ELLs who are highly proficient in the English language (ELL Level 1) may be able to express their subject-specific ideas and understanding by drawing on everyday, non-scientific language. Conversely learners’ with lesser English language proficiency may not. However, as students progress through the secondary phase of education, the spontaneous production of scientific discourse ‘using *precise scientific language*’ becomes compulsory (Department for Education and Skills, 2002: 9) and this may impact more significantly on all ELLs unless they are offered specific support to address these language needs.

After accounting for the other variables, not only did presence or absence of ‘*visuals*’ in assessment questions not appear to differentially affect ELLs’ and ENSs’ performance through an interaction, it also had no overall statistically significant effect on question performance. This finding was unexpected and possible reasons are discussed further in Section 7.

4.3 | Additional Inference: Traits that affect performance but do not interact with language ability

All students were less likely to respond correctly to questions on the conceptually more difficult topics. Compared to a question on a one beaker topic, the log-odds of correctly answering an equivalent question on a two beaker topic were 0.08, with a 95% CI of (-0.77, 0.93); a three beaker topic -1.63 (-2.62, -0.64). There was not therefore much difference between the 1 and 2

beaker topics (low and moderately difficult) but topics rated as 3 beakers (most difficult) were considerably less likely to be answered correctly. More interestingly, we saw no evidence of an interaction between topic difficulty and language proficiency. English native speaking learners seem to be finding more advanced Science topics conceptually just as difficult as ELLs. This supports a view that ELLs do not inherently find more advanced scientific concepts any more difficult than their ENS peers but rather that it is other factors which hinder them in expressing their ability.

Finally, while there was little/no difference between the medium and low-density ELL population schools, those students who came from schools with a high density of ELLs did not do as well as their equivalent (age and language ability) peer in a lower density school (see Supplementary Material, 1). This finding must however be interpreted with care as school ELL density may simply be a proxy for catchment area (e.g. lower socio-economic status of the student population).

5 | DISCUSSION

Our findings provoke further discussion on the validity and equity of current assessment techniques for ELLs, and provide insight into wider linguistic demands in assessment for ELLs. We consider both of these in light of our analysis and using specific illustrations taken from individual learner scripts described further in Section 6.

5.1. | Validity of assessment methods and performance outcomes for ELLs

The very strong evidence we found that specific traits of Science assessment questions influence the size of the performance gap between ELLs and ENSs suggests a wider discussion on the equitable nature of assessment methods for ELLs, and specifically provision of possible alternatives to existing approaches, is critical. In standardized, high stakes summative assessment tasks (the focus of this paper), learners are typically asked to complete largely decontextualized, factual-knowledge demonstration tasks¹. However, it is crucially important for ELLs to be exposed in their classrooms “with robust opportunities to learn” (Schlepelgrell 2012: 416) through formative alternative assessment methods. These include ‘performance assessment’, ‘project-based assessment’ and informal dialogic assessment during ‘inquiry-based Science instruction’ (see for example the work of Maton, 2013). These alternatives aim to allow students to complete tasks via a varied range of performance methods (demonstration, discussion, modeling, reasoning, drawing); with a wide range of language skills (speaking, reading, listening and writing) in highly contextualized settings (laboratory experiment); and collaboratively rather than individually. It has been suggested that such formative alternative approaches allow both ELL and ENS students to perform better (Rivard, 2004; Shaw, Bunch & Geaney, 2010; Smith, Hanks & Erickson, 2017) by providing a broader range of knowledge-demonstration channels to display knowledge and topic-specific expertise. August and Hakuta (1997) and Lyon, Bunch and Shaw (2012) however warn that while some modes of performance assessment may be beneficial to some groups of learners, they may pose additional difficulties to others. Learners with low language proficiency may find it difficult to comprehend the tasks’ instructionally extended and contextually enriched cues, process and respond to their peers’ suggestions, and put forward their own ideas. To address this problem and support progress and performance of ELLs better, Wilmes and Siry (2018) propose using a model whereby learners’ performance can be evaluated drawing not only on their verbal but also on their non-verbal (embodied) modes of interaction. The authors call this model an ‘interaction ritual analysis’.

The consistent differences we observed between the performance of ELLs and ENSs also lead us to consider the wider validity of rating procedures when it comes to evaluating ELLs' performance. Specifically, what is a valid and reliable approach to interpret and score their work as 'atypical, non-mainstream' learners? This issue has been considered not only in the context of ELLs, but also ethnic minorities, certain social groups, refugees, and learners with disruptive schooling experiences. We provide in Section 6 specific examples of individual responses from ELLs that, while being factually correct in terms of subject-specific knowledge, were unrecognized by the standardized marking scheme. Similar effects have also been identified by Noble, Suarez, Rosebery, O'Connor, Warren and Hudicourt-Barnes (2012); equally Warren and Rosebery (1992), and Hamp-Lyons (1991) with unfavorable scoring of ELLs' written performance being observed due to grammatical, syntactical or lexical errors. Shaw (1997) emphasizes the importance of teacher and assessor training to accurately interpret subject-specific performance of socially, culturally and linguistically diverse groups of learners.

Closely related to the issue of rating deficiencies is assessment equity (Siegel, 2007). Lyon (2013) asserts that this occurs when language and experiences which non-mainstream learners bring from their home and personal cultural environments are valued and respected, and where they do not put learners at a disadvantage in demonstrating knowledge. Evidence that ELLs' performance can be treated non-equitably is widespread (Solano-Flores & Trumbull, 2003; Kopriva, 2008). Such mis-assessment of students, based on inequitable teaching and assessment practices, could be considered as part of the 'educational debt' that we owe learners who suffer from achievement gaps (Ladson-Billings, 2006). Again, in Section 6, we present a specific example of such a task in our study that required learners to identify features of a penguin – an animal which is potentially completely unfamiliar, or only partly familiar, to some groups of nonmainstream learners.

5.2. | Linguistic demands of assessment and ELLs' performance outcomes

Our analysis clearly identified the critical importance of language in determining scientific assessment outcomes. ELLs performed consistently more poorly than their ENS peers for all question traits, but were particularly disadvantaged if a question aimed to target scientific vocabulary. Some researchers argue learners should be permitted to demonstrate content-specific knowledge using the linguistic means they feel most comfortable with, be it scientific vocabulary, everyday vocabulary or a combination of both (Brown & Spang, 2007; Lyon, Bunch & Shaw, 2012; Schoerning, Hand, Shelley & Therrien, 2015). The proposed benefits of such a flexible language approach include giving learners greater agency and presence in the classroom by making them feel more able to participate freely and think divergently (Schoerning, Hand, Shelley & Therrien, 2015). Supporters also argue that it allows learners to develop fundamental understanding of scientific ideas and phenomena prior to being asked to operate with them using technical scientific language (Brown & Spang, 2007). A similar argument can be made for formal assessment settings. Expecting learners to produce technical vocabulary, process subject-specific discourse and effectively observe conventions of academic language when they are not yet ready for it, may create space for inaccurate judgments about their *actual*, as opposed to *demonstrated*, knowledge and performance. This needs to be taken into serious consideration when assessing performance of ELLs using assessment instruments developed for mainstream ENS learners.

Performance for all learners (both ELL and ENS) was also significantly reduced if a question required active language production. Duran, Dugan and Weffer (1998: 315) propose that it is not knowledge of scientific vocabulary per se that makes a learner successful in learning Science, but rather the ability to relate, interpret and linguistically assess scientific ideas in a range of semiotic forms. Under this paradigm, teachers should encourage learners to

produce and actively use language to express their ideas and facilitate understanding. Despite this, Lemke (1990) observed that while teachers frequently used target language patterns, learners had very limited opportunities to use these patterns themselves in their own speech. This is common practice where teachers are unaware of the importance of active language instruction and practice (Lyon, 2013; Siegel & Wissehr, 2011; Wong-Fillmore, 2007). Our study showed that ELLs who had lower levels of English competency ('OK' at best) were those most significantly affected if a question required active language production. Thus, encouraging production of Science language becomes even more important for ELLs with lower language competency who may require extra instruction and practice not only in the subject-matter but also in the linguistic means to express it (Wolf & Farnsworth, 2014).

It is important to recognize that the term 'academic language production' should not be seen as synonymous with 'active language production'; although 'active language production' may encompass production of academic language. Specifically, in Science it is common to observe learners completing language production tasks that fall into two language categories. The first, more often seen in traditional examination settings and especially at secondary and tertiary educational phases, requires production of highly scientific academic language. However, the second, more commonly seen at primary level and during class learning rather than standardized assessment, encourages learners to use multiple and diverse types of language, including everyday language, to communicate their understanding of scientific ideas. Recognizing and accepting these diverse linguistic practices in both instructional and assessment settings as valid and acceptable for ELLs, particularly for those with lower levels of English language proficiency, at all educational phases would allow educators to differentiate better between the assessment of scientific language itself from the assessment of scientific ideas. We exemplify this point further in the following section.

6 | IMPLICATIONS

In this section we suggest several implications for practice in assessment of ELLs, and linked implications for teaching. We illustrate these using individual learner responses taken from our study. Some of these practices may also be beneficial for educating ENS learners as they relate to subject- rather than language- specific matters.

Firstly, since ELLs perform less well than their ENS peers, particularly when tested on scientific vocabulary or required to actively produce language, a requirement that has recently been added to the goals of some Science standards (e.g. NGSS), they might have particular difficulties demonstrating their knowledge using these means. We therefore recommend *allowing flexibility in the choice of language (scientific/academic versus non-scientific/non-academic) for assessment/monitoring and teaching purposes of ELLs with lower levels of English language proficiency during a transitional and/or catching up period*. Permitting these learners to draw on everyday, subject non-specific/non-academic language may allow them to better express their conceptual understanding and knowledge. The left hand side of Image 1 presents such an example taken from an ELL student where, despite using casual language, we can see the learner has the subject-specific knowledge.

Image 1: to be inserted here

In our experience, during routine teaching and learning many teachers would have accepted this answer as correct. However, in a formal assessment setting that requires expression of understanding using more advanced language the student would not have scored marks (see current SATs marking scheme, Image 2, specifying no credit should be given for a 'fluffy tummy' answer).

Image 2: to be inserted here

Secondly, in the current educational climate in the UK, production of academic and scientific language, as opposed to everyday language, is one of the core requirements of successful performance at later stages of schooling. In this context it therefore becomes key that, as learners become more familiar with scientific and academic discourse (via teaching instruction), they get actively stretched to *perform* tasks using these types of language. In our study, all three groups performed more poorly when responses to questions required active production of scientific or everyday language used in a specialist science sense to convey scientific ideas. Hence our next recommendation is: *in educational contexts similar to the UK start teaching and eliciting academic and scientific language from learners actively at lower (primary/elementary) stages of schooling in order to prepare learners better for later stages of schooling (secondary phase)*. Furthermore, many learners in our study, both ELLs and ENSs, did not know such subject-specific terminology as: absorb, amount/of, attract, beaker, canine, molar, decay, condense, evaporate/evaporation, feature, grow/th, nutrition, producer, property, reproduce/reproduction, separate, type, vapour, water cycle'. The right hand side of Image 1 above shows an ELL who has ringed the words they did not know. This lack of understanding of key instructional elements made it virtually impossible for them to perform the task illustrating that heavily concentrated technical terms can cause significant comprehension challenges, in agreement with Fang (2006). This fits with the “cumulative knowledge-building” approach described by Maton (2013, p. 2) that aims to enable teachers to work with students in unpacking abstract scientific terms and concepts to develop “more grounded and less condensed meanings” (p. 15), thus enhancing students’ subject-specific understandings as well as their scientific literacy.

The particular difficulty shown by ELLs in responding to questions focusing on scientific vocabulary indicates that these learners may misuse subject-specific vocabulary while still knowing the underlying scientific facts. Image 3 presents three such examples. The first two are the work of ELLs (Punjabi and Urdu speakers) and the third is the work of an ENS learner.

Image 3: to be inserted here

This question, on the labelling of the plant parts, actually shows that it is not just ELLs, but also ENS learners, who can struggle when required to produce highly subject-specific terminology. Also clear here is the need to consider the solutions to the labelling and naming tasks together. If we acknowledge their incorrect labelling, it is clear that all three learners have the knowledge intended to be elicited on the second part naming the parts through which water must pass. However, if these two tasks had been considered in isolation, such understanding would not have been visible. We thus suggest: *when assessing/interpreting learners’ subject specific performance give them multiple opportunities to demonstrate their knowledge on one and the same phenomenon using various means (such as: writing, drawing, labeling, speaking, discussing, performing) and methods (such as: completing combined and multi-level tasks) and consider this performance as a whole allowing ‘parts of the jigsaw’ to fit together*.

Finally, in agreement with Fung and Yip (2014), due to the consistent performance gap between ELLs and ENSs we see across all questions traits and the increase in that gap for those with lower English language ability, we suggest that *ELLs who have limited proficiency in English but who are literate in their first language may be permitted to use their first language for developing and demonstrating their subject-specific knowledge in assessment scenarios*. An example of the potential importance of this can be seen in image 4, where a Hungarian speaker has correctly used their native language to label the plant’s root. Equally, the two ELL speakers

of Image 3 might not have demonstrated vocabulary confusion issues had they been allowed, or rather encouraged, to use their first languages.

Image 4: to be inserted here

7 | LIMITATIONS

Our study has several potential limitations. Firstly, as discussed earlier, school ELL density may be a proxy for the socio-economic status of the student population meaning its interpretation must be treated with care. Secondly, the specific '*first language*' of the ELL may also be an important variable influencing assessment performance. ELLs with different first languages may struggle with different aspects of assessment and require different support. Due to the large number of differing native languages of the ELLs in our study, and the resultant small sample sizes for each, we were unable to perform a rigorous statistical analysis of this but suggest it as a potential topic for further study. Our sample only comprises learners from the UK. It is possible that the needs of learners in other countries may be different. We equally suggest this as an important area for future study. Finally, we note that use of the correct subject-specific, formal vocabulary is a necessary aspect of Science if one wishes to communicate ideas clearly and precisely at a high level. Its inclusion in assessment cannot therefore be removed, the question for educators however is at what point of study its specific assessment should be introduced and when more flexibility should be permitted.

We also recognize that our assessment design was not able to consider all accommodations which may potentially interact with language. For example, the provision of extra time; or permitting students to take assessments written in, and equally respond in, their first language. Importantly, in this regard, while we find no evidence in our study that the presence of illustrations affects performance, it may be that, as suggested by Solano-Flores, Wang, Kachchaf, Soltero-Gonzalez, & Nguyen-Le (2014) this is dependent upon the cultural group under consideration and the specific design of the illustration (see Lohse, Biolsi, Walker & Rueter, 1994 for a potential approach to classification of visuals). Alternatively, it may be that, as in the case of multilingual assessments, students need instruction and formative practice in the classroom to make best use of visuals when they are provided in an assessment. We therefore suggest that the effect of different features of the presented illustration, the amount of practice students have in their use, and their interaction with learner characteristics are important areas for future study. The specific nature of language and linguistic structure in the assessment rubrics and prompts would also be a valuable area for further work.

8 | CONCLUSION

Our study provides strong evidence that language proficiency has an important influence on a learners' ability to answer scientific assessment questions. However, the impact of language proficiency varies significantly according to question trait suggesting the potential to begin to mitigate for language through appropriate assessment design and targeted teaching support. This has important consequences for the design and construction of not just more equitable summative assessment, but also ties closely to the promise of more effective formative assessment in the classroom by enabling the identification of the specific areas of Science that ELLs find more difficult and where more precise and individualized support may be needed.

ELLs did not do as well as their ENS peers for all question types. While the main difference in performance was seen with ELLs who were less proficient in English, even ELLs who spoke English very well performed more poorly than their ENS peers. The greatest detrimental effects on the performance of ELLs, relative to ENSs, were seen on tasks that aimed

to assess formal scientific vocabulary; and/or if the response required active production of language. Here ELLs were particularly disadvantaged compared to their ENS peers. Conversely, assessment questions that targeted scientific fact or research understanding (at least amongst ELLs with good English proficiency), or that did not require the active production of language showed a much reduced performance gap. These conclusions lead us to suggest that ELLs may often possess the intended underlying scientific understanding but lack the required vocabulary and language skills to demonstrate it appropriately during assessment. This argument is supported by our findings that the gap in performance between ELLs and their ENS peers is not significantly altered by the difficulty of the topic under assessment suggesting differences in achievement are not influenced by conceptual difficulty.

We also see that it is not ELLs alone who experience difficulties in acquiring scientific content (i.e. the subject-matter itself) and scientific discourse (i.e. the language of Science, as part of learners' cognitive academic language proficiency); many ENS children also find these tasks difficult. The more difficult the tasks were conceptually and/or the less flexibility they allowed in the language-response format the more all groups of learners struggled.

In conclusion, the changing nature of assessment in relation to emerging frameworks (e.g. NGSS) brings about new language demands on students. This can be particularly demanding for ELLs. Teachers and assessors need to be responsive to new practices and there is an important role for education professionals to promote discipline-specific learning through appropriate, formative and equitable pedagogies. This includes recognizing the multiple educational, linguistic and socio-cultural dimensions that ELLs bring into the classroom. Moreover, the notion of educational debt needs further consideration in teaching, learning and assessment processes.

Science requires active language production to successfully communicate ideas. However many traditional science assessment questions equate such a need for active language with production of highly scientific language only. An alternative, and potentially more equitable, approach which assessors could consider would be to reframe the requirement for active language and encourage more varied types of language production in assessment tasks, clearly separating assessment of scientific language and science ideas. Such a change would require a significant shift from many current assessment practices, as well as careful integration with later stages of study where precise use of specific scientific language does become critical. Whatever the implications for assessment, for teachers in the classroom, we suggest a particular focus is needed on developing, through suitable pedagogy and formative assessment, ELLs' skills in producing both active language and their ability to use/recognize formal scientific vocabulary. In this way, such learners should have more equitable opportunities to access the content of their respective national curricula and to demonstrate their knowledge in ways that enhance their performance rather than restrict it.

ACKNOWLEDGMENTS

We would like to thank our anonymous reviewers for their detailed and insightful comments which have undoubtedly helped us to develop our thinking and arguments further. We are also immensely grateful for the support for this research provided by a Local Authority as well as the local schools, teachers and pupils who participated in this research.

ENDNOTES

¹ It is of course possible to use in class summative assessments formatively whereby teachers and their learners discuss the responses to questions as the means to support learners in both their ability to make meaning in Science and develop their subject knowledge.

REFERENCES

- Author1 (2015)
- Author1 and Author2 (2016)
- Author2 and Author1 (2011)
- Abedi, J. (2002). Assessment and accommodations of English language: Issues, concerns, and recommendations. *Journal of School Improvement*, 3(1), 83-89.
- Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, 33(1), 4-14.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Anstrom, K., DiCerbo, P., Butler, F. A., Katz, A., Millet, J., & Rivera, C. (2010). *A review of the literature on academic English: Implications for K-12 English language learners*. Arlington, VA: George Washington University Center for Equity and Excellence in Education.
- Ardac, D., & Akaygun, S. (2004). Effectiveness of multimedia-based instruction that emphasizes molecular representations on students' understanding of chemical change. *Journal of Research in Science Teaching*, 41, 317-337.
- August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: A research agenda*. Washington, DC: National Academy Press.
- Barnett, J. (1992). Language in the science classroom: some issues for teachers. *Australian Science Teachers Journal*, 38(4), 8-13.
- Bates, D., Maechler, M., Bolker, B. & Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi: 10.18637/jss.v067.i01
- Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen J.R., Stevens M.H., & White, J.S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3), 127-135, doi: 10.1016/j.tree.2008.10.008
- Bravo, M. A., Mosqueda, E., Solis, J. L., & Stoddard, T. (2014). Possibilities and limits of integrating Science and diversity education in preservice elementary teacher preparation. *Journal of Science in Teacher Education*, 25, 601-609.
- Brown, B. A. & Spang, E. (2007). Double talk: Synthesizing everyday and science language in the classroom. *Science Education*, 92, 708-732. DOI: 10.1002/sce.20251
- Buck Bracey, Z.E. (2017). Students from non-dominant linguistic backgrounds making sense of cosmology visualizations. *Journal of Research In Science Teaching*, 54(1), 29-57.
- Bunyi, G. (1999). Rethinking the place of African indigenous languages in African education. *International Journal of Educational Development*, 19(4-5), 337-350.
- Buxton, C., Alleksaht-Snyder, M., Aghasaleh, R., Kayumova, S., Kim, S., Choi, Y. and Cohen, A. (2014). Potential benefits of bilingual constructed response science assessments for understanding bilingual learners' emergent use of language of scientific investigation practices. *Double Helix*, (2), 1, 1-21.
- Cleghorn, A., Merritt, M., & Abagi, J.O. (1989). Language policy and Science instruction in Kenyan primary schools. *Comparative Education Review*, 33(1), 21-39.

- Cromley, J. G., Weisberg, S. M., Dai, T., Newcombe, N. S., Schunn, C. D., Massey, C. & Merlino, F. J. (2016) Improving Middle School Science Learning Using Diagrammatic Reasoning. *Science Education*, 100(6), 1185-1213. DOI 10.1002/sce.21241
- Curtis, S., & Millar, R. (1988). Language and conceptual understanding in science: A comparison of English and Asian language speaking children. *Research in Science and Technological Education*, 6(1), 61–77.
- Dempster, E. R. & Reddy, V. (2007). Item readability and science achievement in TIMSS 2003 in South Africa. *Science Education*, 91, 906-925. DOI: 10.1002/sce.20225
- Department for Education (2016). *Schools, pupils, and their characteristics. National Statistics*. London: Department for Education (accessed on 12 December 2017).
- Department for Education and Skills. (2002). *Key Stage 3 National Strategy: Access and engagement in science. Teaching pupils for whom English is an additional language*. DfES Publications. DfES 0610/2002
- Duran, B. J., Dugan, T. & Weffer, R. (1998). Language minority students in high school: The role of language in learning biology concepts. *Science Education*, 82, 311-341.
- Evans, M., Schneider, C., Arnot, M., Fisher, L., Forbes, K., Hu, M. & Liu, Y. (2016). *Language development and school achievement: Opportunities and challenges in the education of EAL students*. University of Cambridge, Anglia Ruskin University and The Bell Educational Trust Limited (Executive Summary).
- Fang, Z. (2006) The language demands of Science reading in middle school, *International Journal of Science Education*, 28(5), 491-520, DOI: 10.1080/09500690500339092
- Flores, N. and Rosa, J. (2015). Undoing appropriateness: raciolinguistic ideologies and language diversity in education. *Harvard Educational Review*, (85), 2, 149-171.
- Frantz, R.S., Bailey, A.L., Starr, L. & Perea, L. (2014) Measuring Academic Language Proficiency in School-Age English Language Proficiency Assessments Under New College and Career Readiness Standards in the United States, *Language Assessment Quarterly*, 11:4, 432-457, DOI: 10.1080/15434303.2014.959123
- Fung, D. & Yip, V. (2014). The effects of the medium of instruction in certificate-level physics on achievement and motivation to learn. *Journal of Research In Science Teaching*, 51(10), 1219–1245.
- Garcia, E.E. (1997). Multilingualism in U.S. schools: Treating language as a resource for instruction and parent involvement. *Early Child Development and Care*, 127–128, 141–155.
- Gee, J. P. (2008). What is academic language? In A. S. Rosebery & B. Warren (Eds.), *Teaching science to English language learners: Building on students' strength*. (pp. 57–69). Arlington: National Science Teachers Association.
- Hammond, J. (2014). An Australian perspective on standards-based education, teacher knowledge, and students of English as an additional language, *TESOL Quarterly*, 48(3) 507-532. doi: 10.1002/tesq.173
- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–278). Norwood, NJ: Ablex.
- Honeycutt Swanson, L., Bianchini, J. A., & Lee, J. S. (2014). Engaging in argument and communicating information: A case study of English language learners and their Science

- teacher in an urban high school. *Journal of Research in Science Teaching*, 51(1), 31–64. <https://doi.org/10.1002/tea.21124>
- Kiboss, J. K., Ndirangu, M., & Wekesa, E. W. (2004). Effectiveness of a computer-mediated simulations program in school biology on pupils' learning outcomes in cell theory. *Journal of Science Education and Technology*, 13, 207–213.
- Kieffer, M. J., Lesaux, N., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 29(3), 1168–1201. DOI: 10.3102/003465430933249
- Kopriva, R.J. (2008). Improving testing for English language learners. New York: Routledge.
- Ladson-Billings, G. J. (2006). From the achievement gap to the education debt: understanding achievement in U.S. schools. *Educational Researcher*, 35(7), 3-12.
- Lara-Alecio, R., Tong, F., Irby, B.J., Guerrero, C., Huerta, M., & Fan, Y. (2012) The effect of an instructional intervention on middle school English learners' science and English reading achievement. *Journal of Research In Science Teaching*, 49(8), 987–1011.
- Lee, O. (2004). Teacher change in beliefs and practices in science and literacy instruction. *Journal of Research in Science Teaching*, 41(1), 65–93.
- Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research*, 75(4), 491–521. <https://doi.org/10.3102/0034654075004491>
- Lee, O., Quinn, H. and Valdes, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English language arts and mathematics. *Educational Researcher*, (42), 4, 223–233. DOI: 10.3102/0013189X13480524
- Lemke, J. L. (1990) *Talking Science: Language, Learning, and Values*. New York: Ablex Publishing Corporation.
- Lohse, G. L., Biolsi, K., Walker, N & Rueter, H. H. (1994). *A Classification of Visual Representations*. Commun. ACM. 37. 36-49. 10.1145/198366.198376.
- Lyon, E. G. (2013), Learning to assess science in linguistically diverse classrooms: Tracking growth in secondary science preservice teachers' assessment expertise. *Science Education*, 97: 442-467. doi:10.1002/sce.21059
- Lyon, E. G., Bunch, G. C., & Shaw, J. M. (2012). Navigating the language demands of an inquiry-based science performance assessment: Classroom challenges and opportunities for English learners. *Science Education*, 96, 631–651. <https://doi.org/10.1002/sce.21008>
- Maerten-Rivera, J., Myers, N., Lee, O., & Penfield, R. (2010). Student and school predictors of high-stakes assessment in Science. *Science Education*, 94, 937–962.
- Martin, J.R. (1993) Literacy in Science: learning to handle text as technology. In Martin, J.R. and Halliday, M.A.K. (1993) *Writing Science: Literacy and Discursive Power*. Pittsburgh: University of Pittsburgh Press.
- Maton, K. (2013). Making semantic waves: a key to cumulative knowledge-building. *Linguistics and Education*, 24, 8-22.
- McCloskey, M. (2002). President's message: No child left behind. *TESOL Matters*, 12(4). Retrieved from <http://www.tesol.org/pubs/articles/2002/tm12-4-04.html>

- McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall. ISBN: 9780412317606.
- Moje, E.B. (2015). Doing and teaching disciplinary literacy with adolescent learners: A social and cultural enterprise. *Harvard Educational Review*, 85(2), 254–278. <https://doi.org/10.17763/0017-8055.85.2.254>
- Murphy, V., & Unthiah, A. (2015). *A systematic review of intervention research examining English language and literacy development in children with English as an Additional Language (EAL)*. University of Oxford: Department of Education.
- National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.
- Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). 'I never thought of it as freezing': How students answer questions on large-scale science tests and what they know about science. *Journal of research in science teaching*. 49(6), 778-803. DOI: 10.1002/tea.21026
- Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (No. NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
- Pitoniak, M.J., Young, J.W., Martiniello, M., King, T.C., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English language learners*. Princeton, NJ: Educational Testing Service.
- Proposition 227. (1998). *English language in public schools*. Initiative Statute. Sacramento, CA: Attorney General.
- Qualifications and Curriculum Authority (2003-2011). *Science tests, Key Stage 2, Levels 3-5*. QCA Publications, Sudbury: Suffolk.
- R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. URL <https://www.R-project.org/>.
- Rainey, E., L. Maher, B., Coupland, D. Franchi, R. & Moje, E. (2017). But What Does It Look Like? Illustrations of Disciplinary Literacy Teaching in Two Content Areas. *Journal of Adolescent & Adult Literacy*. 61. 371-379. 10.1002/jaal.669.
- Rea-Dickins, P., Khamis, Z., & Olivero, F. (2013). Does English-medium instruction and examining lead to social and economic advantage? Promises and threats: a Sub-Saharan case study. In E. Erling & P. Seargeant (Eds.) *English and International Development*. Avon: Multilingual Matters Ltd.
- Rivard, L. (2004). Are language-based activities in science effective for all students, including low achievers? *Science Education*, 88, 420-442. DOI: 10.1002/sce.10114
- Robinson, P.J. (2005). Teaching key vocabulary in geography and science classrooms: An analysis of teachers' practice with particular reference to EAL Pupils' Learning, *Language and Education*, 19(5), 428-445, DOI: 10.1080/09500780508668695
- Routitsky, A., & Turner, R. (2003). *Item format types and their influence on cross-national comparisons of student performance*. Chicago, USA: Presentation given to the Annual Meeting of the American Educational Research Association (AERA) in Chicago, USA.

- Ryu, M. (2015). Positionings of racial, ethnic, and linguistic minority students in high school biology class: implications for science education in diverse classrooms. *Journal of Research In Science Teaching*, 52(3), 347–370.
- Schleppegrell, M. (2012). Academic language in teaching and learning: Introduction to the special issue. *The elementary School Journal*, (112), 3, 409-418.
- Schoerning, E. Hand, B. Shelley, M. & Therrien, W. (2015) Language, access, and power in the elementary science classroom. *Science Education*, 99(2), 238-259. DOI: 10.1002/sce.21154.
- Shanahan, T., & Shea, L. (2012). Incorporating English language teaching through Science for K-12 teachers. *Journal of Science Teacher Education*, 23, 407–428.
- Shaw, J. M. (1997). Threats to the validity of science performance assessments for English language learners. *Journal of Research in Science Teaching*, 34(7), 721-743.
- Shaw, J. M., Bunch, G. C., & Geaney, E. R. (2010). Analysing language demands facing English learners on science performance assessment: The SALD framework. *Journal of Research in Science Teaching*, 47(8), 909-928. DOI: 10.1002/tea.20364.
- Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life Science items. *Journal of Research in Science Teaching*, 44(6), 864–881.
- Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education*, 22(4), 371 – 391.
- Smith, L. K., Hanks, J. H. & Erickson, L. B. (2017). Secondary biology textbooks and national standards for English learners. *Science Education*, 101(2), 302-332. DOI: 10.1002/sce.21265
- Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3-13.
- Solano-Flores, G., Wang, C., Kachchaf, R., Soltero-Gonzalez, L., & Nguyen-Le, K. (2014). Developing testing accommodations for English language learners: Illustrations as visual supports for item accessibility. *Educational Assessment*, 19(4), 267-283, DOI: 10.1080/10627197.2014.964116
- Solorzano, R.W. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research*, 78(2), 260–329.
- Strand, S. Malmberg, L. & Hall, J. (2015). *English as an Additional Language (EAL) and educational achievement in England: An analysis of the National Pupil Database*. University of Oxford: Department of Education.
- U.S. Department of Education, National Center for Education Statistics. (2017). The Condition of Education (2017-144), English Language Learners in Public Schools.
- United Nations (1989) *Convention on the Rights of the Child*. Retrieved from <https://www.unicef.org.uk/what-we-do/un-convention-child-rights/>
- Unsworth, L. & Cleirigh, C. (2009) Multimodality and reading: the construction of meaning through image-text interaction. In Jewitt, C. (Ed) *The Routledge Handbook of Multimodal Analysis*. New York: Routledge.

- Valdés, G. (2004) Between Support and Marginalisation: The Development of Academic Language in Linguistic Minority Children, *International Journal of Bilingual Education and Bilingualism*, 7:2-3, 102-132, DOI: 10.1080/13670050408667804
- Wallace, C. S. (2004), Framing new research in science literacy and language use: Authenticity, multiple discourses, and the “Third Space”. *Science Education*, 88: 901-914. doi:10.1002/sce.20024
- Wang, C. (2012). *The use of illustrations in large-scale science assessment: A comparative study* (Unpublished doctoral dissertation). University of Colorado Boulder
- Warren, B., & Rosebery, A. (1992). Science education as sense-making practice: Implications for assessment. In *Focus on Evaluation and Measurement: Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues* (Vol. 2, pp. 273–304).
- Wellington, J. & Osborne, J. (2001). *Language and literacy in science education*. Buckingham: Open University Press.
- Wilmes, S. E. D. & Siry, C. (2018). Interaction rituals and inquiry-based science instruction: Analysis of student participation in small-group investigations in a multilingual classroom. *Science Education*, 102, 1107-1128. DOI: 10.102/sce.21462
- Wolf, M. K., & Farnsworth, T. (2014). English language proficiency assessments as an exit criterion for English learners. In Kunnan, A. (Ed.), *The companion to language assessment* (pp. 303–317). New York: Wiley-Blackwell.
- Wong-Fillmore, L. (2007). English learners and mathematics learning: Language issues to consider. *Assessing Mathematical Proficiency*, 53, 333 – 344.
- Yip, D. Y. (2004). Questioning skills for conceptual change in science instruction, *Journal of Biological Education*, 38(2), 76-83, DOI: 10.1080/00219266.2004.9655905

TABLES

Table 1: An overview of the compulsory education system in the UK

Phase	Age	School Year	Stage	Examinations
Foundation	4-5	Reception	N/A	
Primary	5-6	Year 1	Key Stage 1	SATs – Standard Attainment Tests are used to evaluate children’s educational progress at the end of Years 2, 6 and 9
	6-7	Year 2		
	7-8	Year 3	Key Stage 2	
	8-9	Year 4		
	9-10	Year 5		
	10-11	Year 6		SATs (Science – sampling only)
Secondary	11-12	Year 7	Key Stage 3	
	12-13	Year 8		
	13-14	Year 9		SATs
	14-15	Year 10		

	15-16	Year 11	Key Stage 4	GCSEs - qualifications in specific subjects, as part of the General Certificate of Education, at a level below Advanced level.
Sixth Form / College	16-17	Year 12	N/A	A Levels/IB & NVQs/BTECs – qualifications in specific subjects, as part of the General Certificate of Education, at Advanced level.
	17-18	Year 13		

Table 2: Distribution of learner cases by school and year group

Year Group / Age	School 1	School 2	School 3	School 4	School 5
Year 3 (7-8 years)	31	28	30	29	25
Year 4 (8-9 years)	0	24	26	20	27
Year 5 (9-10 years)	0	23	26	24	17
Year 6 (10-11 years)	29	20	25	15	21

Table 3: Number of students assessed on each topic by language proficiency class

Topic No.	Total number of students per language proficiency class					
	223	100%	168	100%	49	100%
	Number of students attempting topics per language proficiency class					
	Native Eng		ELL Level 1		ELL Level 2	
1	114	51%	91	54%	34	69%
2	223	100%	167	99%	48	98%
3	83	37%	82	49%	22	45%
4	83	37%	82	49%	22	45%
5	109	49%	76	45%	15	31%
6	109	49%	76	45%	15	31%
7	68	30%	39	23%	3	6%
8	68	30%	39	23%	3	6%

Table 4: Approximate p-value from LRT

Term	Approximate p-value from LRT
Year Group (3 levels)	1.0×10^{-6}
School ELL Density (3 levels)	7.9×10^{-9}
Topic Difficulty (3 levels)	0.003
Interaction between language proficiency and active language production (3 x 2 levels)	0.014
Interaction between language proficiency and question focus (3 x 4 levels)	0.002

Table 5: Log-odds ratio for various combinations of ‘question focus’ and language proficiency.

	Scientific fact (SF)		Scientific fact and vocabulary (SFV)		Research (R)		[Scientific] Vocabulary (V)	
	For questions requiring active language production							
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Native English	Baseline	N/A	-0.45	0.58	1.86	0.78	0.66	0.70
ELL Level 1	-0.06	0.17	-0.57	0.61	1.54	0.80	-0.27	0.73
ELL Level 2	-1.05	0.27	-0.98	0.72	0.93	0.86	-1.73	0.81

Table 6: Log-odds ratio for various combinations of ‘language production type’ and language proficiency.

	Active language production – language production group		No active language production – language recognition group	
	For questions of “Scientific fact” (SF) assessment type			
	Estimate	Std. Error	Estimate	Std. Error
Native English	Baseline	N/A	0.40	0.35
ELL Level 1	-0.06	0.17	0.30	0.42
ELL Level 2	-1.05	0.27	0.02	0.48

FIGURE LEGENDS

Figure 1: Overview of topics by year group at KS2

Figure 2: Question characteristics by focus, visual, language production and difficulty.

Figure 3: Proportion (and estimated 95% confidence intervals) answering at least partly correctly for the three language proficiency classes on each question.

IMAGE LEGENDS

Image 1: The LHS shows an example of a student expressing understanding using non-scientific language; the RHS possible problems comprehending language of instruction

Image 2: SATs marking scheme guide (excerpt)

Image 3: Demonstration of knowledge drawing on multiple tasks

Image 4: Use of first language in Science

Figure 1: Overview of topics by year group at KS2

Scale of Sampling by Year Group	Year 3		Year 4				Year 5		Year 6	
	Year 6									
Topic	Teeth and eating	Growing plants	Magnets	Habitats and Food chain		Separating solids and liquids	Changing state		Changing sounds	Changing circuits
Taught in	Y3 topic	Y3 topic	Y3 topic	Y4 topic		Y4 topic	Y5 topic		Y5 topic	Y6 topic

Figure 2: Question characteristics by focus, visual, language production and difficulty

Topic	Teeth and eating				Growing plants				Magnets				Habitats and Food chain				Separating solids and liquids				Changing state				Changing sounds				Changing circuits			
Difficulty																																
Question No.	1a	1b	1c	1d	2a	2b	2c	2d	3a	3b	3c	3d	4a	4b	4c	4d	5a	5b	5c	5d	6a	6b	6c	6d	7a	7b	7c	7d	8a	8b	8c	8d
Focus	R	SF	SF	SF	SF	V	SF	SF V	SF	SF	SF	SF	SF	SF	SF V	SF V	SF	SF	SF	SF V	SF	R	SF	SF	SF	SF	SF V	V	SF	SF	V	SF
Visual	✓		✓			✓			✓	✓		✓	✓	✓	✓		✓		✓			✓	✓	✓	✓	✓	✓	✓	✓		✓	
Task type	D	N	E	T/F	N	N	N	N	N	D	E	E	NE	CD	T	E	L	M	DB	E	T	DB	CD	M	N	DB	T	N	D	Y/N	N	N
Language Production	P	A	A	P	A	A	A	A	A	P	A	A	A	P	P	A	P	P	A	A	P	A	P	P	A	A	P	A	P	A	P	A

Difficulty: - Least conceptually demanding topics; - More conceptually demanding topics; - Most conceptually demanding topics
Focus: R – Research procedure; SF – Scientific fact; V – Scientific vocabulary; SF & V - Scientific fact & scientific vocabulary
Visual: ✓ – Visual present; □ – Visual absent
Task type: D – Draw; N – Name; E – Explain; T/F – True/False; L – Label; NE – Name and Explain; CD – Complete Diagram; T – Tick; M – Match; DB – Describe; Y/N – Yes/No
Language production: A – Active language production (i.e. spontaneous language creation); P – Passive language reproduction

Figure 3: Proportion (and estimated 95% confidence intervals) answering at least partly correctly for the three language proficiency classes on each question.

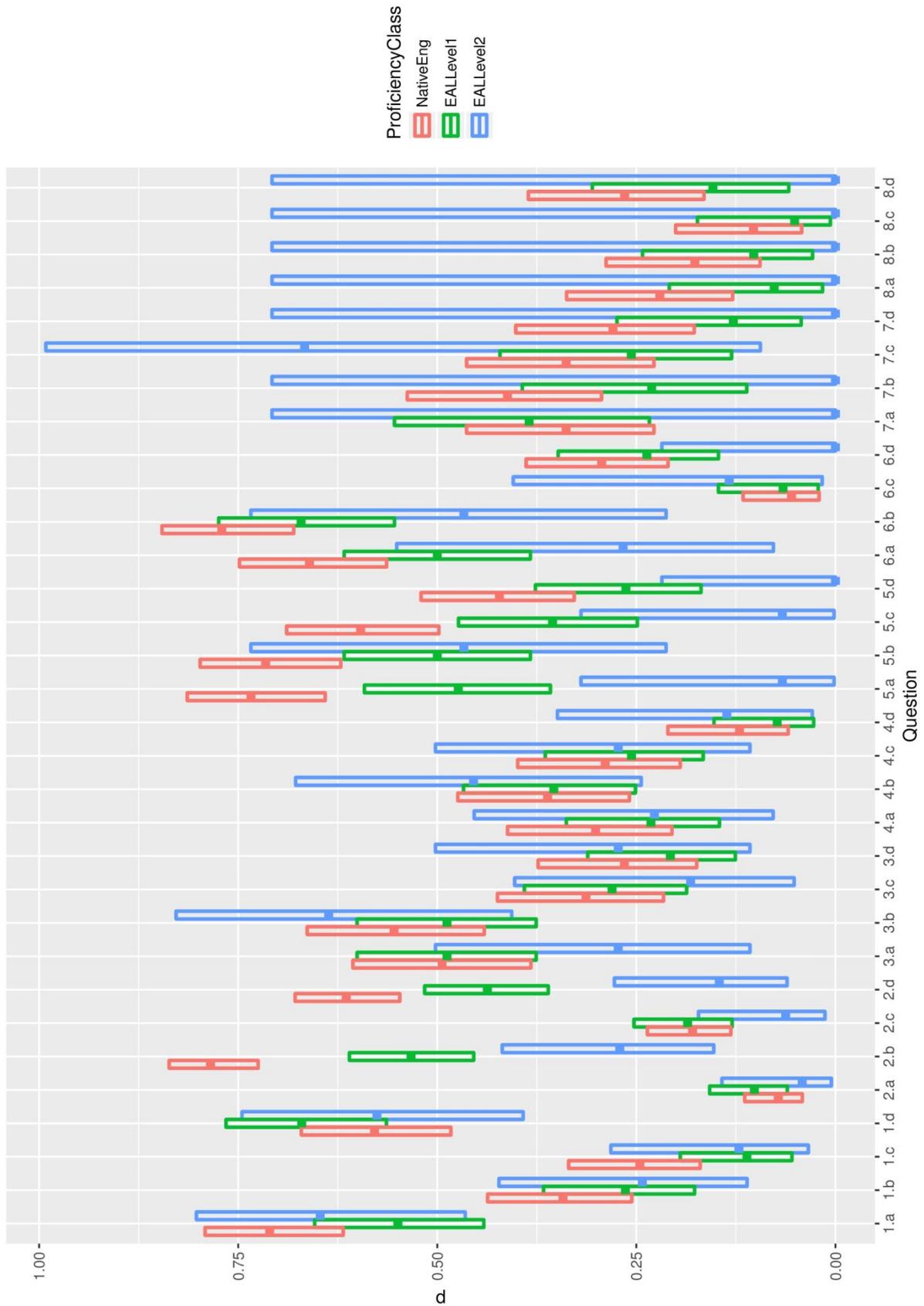


Image 1: The LHS shows an example of a student expressing understanding using non-scientific language; the RHS possible problems comprehending language of instruction

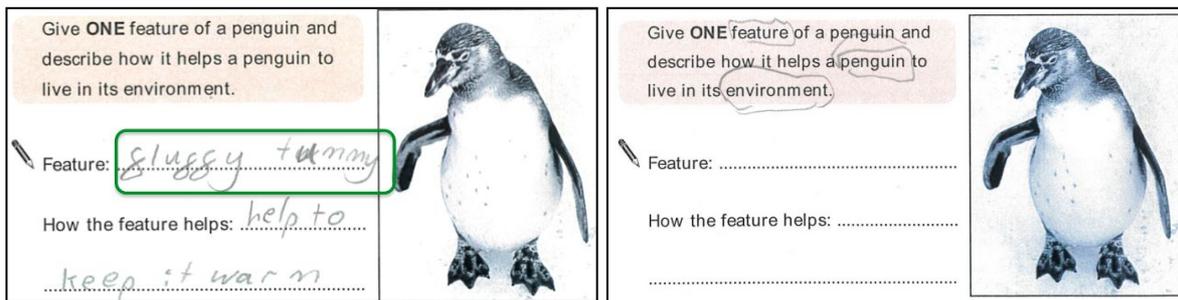


Image 2: SATs marking scheme guide (excerpt)

Mark	Requirements	Allowable answers	Additional guidance
1m	<p>Award ONE mark for a feature of a penguin and a description of how it helps a penguin to live in its environment:</p> <ul style="list-style-type: none"> ■ (thick) feathers it keeps them warm ✓ ■ webbed feet to allow them to swim/to walk on snow ■ streamlined shape for swimming ■ layer of fat/blubber it insulates them ■ a rounded body reduces heat loss/allows it to slide on ice ■ white tummies/a black back animals swimming underneath/above them cannot see them easily ■ flippers they can pull themselves through the water ■ beak to eat/catch fish. 	<p>ONE mark may be awarded for a feature of a penguin's behaviour rather than its body which accurately describes how it may help the penguin to live:</p> <ul style="list-style-type: none"> ■ huddling together keeps them warm. <p>ONE mark may be awarded for a response confusing the penguins' feathers with fur [specific knowledge of penguins' anatomy is not required]:</p> <ul style="list-style-type: none"> ■ fur/hair/thick coat to help keep them warm. ✓ 	<p>Do not give credit for an insufficient response that identifies a feature of a penguin but omits or gives an insufficient explanation of how that feature helps the penguin live in its environment:</p> <ul style="list-style-type: none"> ■ flippers in the water [does not describe how flippers help in water]. <p>Do not give credit for an insufficient response giving a generalised feature of many animals even when an appropriate explanation is given:</p> <ul style="list-style-type: none"> ■ feet it helps them walk/balance ■ coat helps to keep it warm ✗ ■ arms to help swim.

Image 3: Demonstration of knowledge drawing on multiple tasks

Task (with answers)	Learner performance (Punjabi speaker)
<p>The pictures below show different types of flowering plant.</p> <p>Write the THREE missing labels to show the names of the plant parts.</p> <p>Plants absorb rain water from the soil.</p> <p>Name the TWO parts of the plant the water must travel through to get from the soil to the leaves.</p> <p>1. <u>Root(s)</u></p> <p>2. <u>Stem / Stalk</u></p>	<p>The pictures below show different types of flowering plant.</p> <p>Write the THREE missing labels to show the names of the plant parts.</p> <p>Name the TWO parts of the plant the water must travel through to get from the soil to the leaves.</p> <p>1. <u>roots</u></p> <p>2. <u>Veeds</u></p>
Learner performance (Urdu speaker)	Learner performance (English native speaker)
<p>The pictures below show different types of flowering plant.</p> <p>Write the THREE missing labels to show the names of the plant parts.</p> <p>Name the TWO parts of the plant the water must travel through to get from the soil to the leaves.</p> <p>1. <u>Roots</u></p> <p>2. <u>Strips</u></p>	<p>The pictures below show different types of flowering plant.</p> <p>Write the THREE missing labels to show the names of the plant parts.</p> <p>Name the TWO parts of the plant the water must travel through to get from the soil to the leaves.</p> <p>1. <u>Petals</u></p> <p>2. <u>Stalk</u></p>

Image 4: Use of first language in science

Learner performance (Hungarian speaker)

The pictures below show different types of flowering plant.

Write the **THREE** missing labels to show the names of the plant parts.

Plant A

Plant B

flower

stem

leaves

gyökér

gyökér = root

Supplementary Material 1

Table S1: School Profiles

Density of ELL learners		School code	School profile (selected statements) ¹
High 85-100%	96%	Sch1	Almost all pupils are from minority ethnic groups (most recent arrivals are from Eastern Europe - Gypsy/Roma ethnicity). The vast majority of pupils speak English as an additional language. The proportion of pupils supported by the pupil premium ² is higher than national average .
	96%	Sch2	Almost all pupils are from minority ethnic groups (most recent arrivals are from Eastern Europe - Gypsy/Roma ethnicity). The vast majority of pupils speak English as an additional language. The proportion of pupils supported by the pupil premium is higher than national average .
Medium 35-80%	78%	Sch3	The majority of pupils are from minority ethnic groups. A well-above average proportion of them speak English as an additional language. Over 24 different languages are represented in the school. The proportion of pupils supported by the pupil premium funding is well above the national average.
	37%	Sch4	The proportion of pupils from minority ethnic groups and who speak English as an additional language is above average . The proportion of pupils for whom the school receives the pupil premium is significantly below average .
Low 0-30%	17%	Sch5 ³	Most of the pupils are of White British heritage . The proportion of pupils for whom the school receives the pupil premium is below the national average.

¹ Statements are accurate for 2013, the year when the project commenced.

² The pupil premium is an additional government funding for disadvantaged pupils known to be eligible for free school meals and for children who are looked after by the local authority.

³ Sch5 school took part only in the first phase of the research.

Supplementary Material 2: Examples of Assessment Tasks

Image S1: Understanding of research procedures (R)

1(a) Sue wants to find out how four **different** drinks affect teeth.

Egg shell and teeth are made of the same type of material.
 Sue puts the same amount of egg shell in four beakers.
 She puts a **different** drink into each beaker.

Show how much drink Sue must put in each beaker for her test to be fair. Draw a line on beakers B, C and D.

Beaker A has been done for you.

Image S2: Understanding of a scientific fact (SF)

3(c) Nisha moves a different bar magnet towards the magnet on the engine. The magnets do not touch each other.
 The engine moves away from Nisha's magnet.

Explain why the train engine moves away from the bar magnet.

.....

.....

Image S3: Production or recognition of scientific vocabulary (V)

7(d) What is the scientific name for how high or low a note is?

.....

Image S4: Understanding of a scientific fact AND production or recognition of scientific vocabulary (SFV)

5(d) Ahmed mixes salt and water.
 Salt and water cannot be separated with any sieve.

(i) Explain what happens to the salt when he mixes it with water.

.....

Image S5: Name and Explain (N & E) category

4(a) Give **ONE** feature of a penguin and describe how it helps a penguin to live in its environment.



 Feature:

How the feature helps:

.....

Image S6: Describe (DB) category

5(c) Philip needs to clean the fish tank. He takes the fish and the plants out of the fish tank.

The teacher tips the dirty water and gravel from the fish tank into a sieve.

 **Sieve**

Complete the sentences below to show what happens to the gravel and the water when they are separated with the sieve.

 The gravel

The water

Image S7: True/False (T/F) category

1(d) Write **true** or **false** next to each of the statements below.

True or false?

 Children lose their first teeth and grow new teeth.

Human teeth can reproduce.

Image S8: Label (L) category

5(a) Philip's class has some goldfish in a fish tank. The picture below shows the fish tank.

Write **solid**, **liquid** or **gas** to complete each label on the diagram.

One has been done for you.

plastic lid **inside the bubble**

 **solid** (.....) (.....)



water **gravel**

(.....) (.....)

Image S9: Complete Diagram (CD) category

6(c) Write the letters **A-E** on the diagram to show the order of the stages in the water cycle.

One stage is done for you.

Stages of the water cycle

A. water vapour starts to cool down	B. water collects in rivers and lakes
C. rain falls	D. water vapour condenses
E. water evaporates	

Image S10: Tick (T) category

6(a) Rose knows that water and vinegar evaporate.

Tick **ONE** box to show what **evaporation** means.

Evaporation is the change from...

gas to liquid. <input type="checkbox"/>	gas to solid. <input type="checkbox"/>
liquid to solid. <input type="checkbox"/>	liquid to gas. <input type="checkbox"/>

Image S11: Match (M) category

6(d) This is a diagram of the water cycle.

cloud
 hill
 river
 lake

Draw **FOUR** lines to match each letter, A, B, C and D to the correct description of what is happening in the water cycle.

A	Rain falls.
B	Water changes into water vapour.
C	Water vapour changes into water.
D	Water flows into lakes or seas.

Image S12: Yes/No (Y/N) category

8(b) Polly wants the star to shine more brightly.
She has some ideas about how she can do this.

Write **yes** or **no** next to each idea to show if Polly will see the star shine more brightly.

Idea	Will the star shine more brightly? Yes or no?
 add another bulb
add another cell
use longer wires

For Peer Review

Supplementary Material 3: Data Collection

Figure S1: Language background questions

ABOUT YOU:

1. Your gender: Boy  Girl 

2. Is English your first language?

Yes No

3. If English is NOT your first language, how well do you speak it?

Very well 😊 OK 😐 Not very well 😞

4. What language do you speak at home? _____

5. How long have you lived in England (the UK)?

I was born here 1-2 years 3-5 years More than 5 years

Figure S2: Task instructions

INSTRUCTIONS:

 Answer the questions below. Try your best.

 You **CAN** answer the questions in your **first language**.

 If you choose to answer the questions in your first language, please **ALWAYS** try to answer them in English **FIRST**.

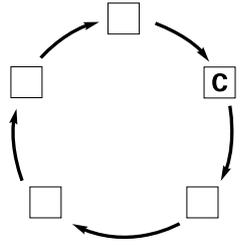
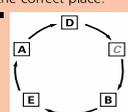
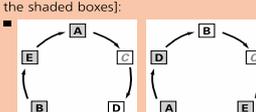
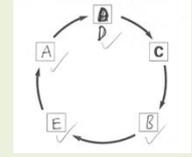
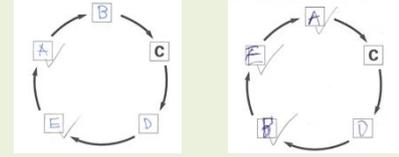
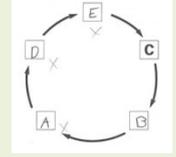
 Underline or circle words that you **DO NOT** understand.

 Try to finish this exercise in **25 minutes**. You can take longer if you need.

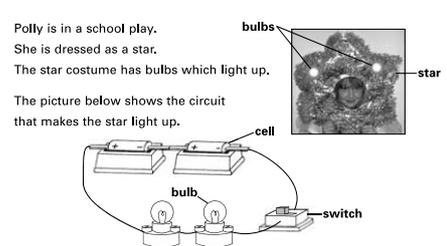
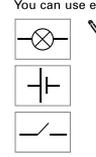
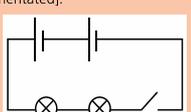
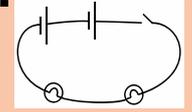
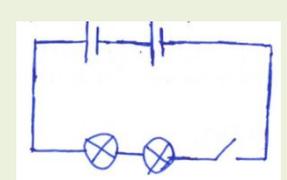
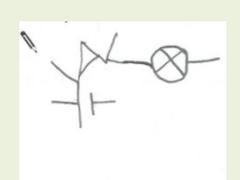
SupplementaryMaterial

4: Examples of learner responses by "correctness" category

Example1: Water cycle

Question	Mark	Requirements	Alawable answers	Additional guidance
<p>Write the letters A-E on the diagram to show the order of the stages in the water cycle.</p> <p>One stage is done for you.</p> <div style="border: 1px solid black; padding: 5px; margin: 10px 0;"> <p style="text-align: center;">Stages of the water cycle</p> <p>A. water vapour starts to cool down B. water collects in rivers and lakes</p> <p>C. rain falls D. water vapour condenses E. water evaporates</p> </div> 	<p>2m</p> <p>or</p> <p>1m</p>	<p>Award TWO marks for all four letters in the correct place:</p>  <p>If you are unable to award two marks, award ONE mark for any two or three letters in the correct place.</p>	<p>ONE mark may be awarded for three letters in the correct order but incorrectly placed on the diagram [the only two possible correct responses indicated by the shaded boxes]:</p> 	
Examples of learner responses				
	Entirely correct (2)	Partly correct (1 mark)	Entirely incorrect (0 marks)	
				

Example 2: Electric circuit

Question	Mark	Requirements	Alawable answers	Additional guidance
<p>Polly is in a school play. She is dressed as a star. The star costume has bulbs which light up. The picture below shows the circuit that makes the star light up.</p>  <p>Draw a circuit diagram of the star's circuit in the space below.</p> <p>Use these symbols in your circuit diagram. You can use each symbol more than once if you need to.</p> 	<p>2m</p> <p>or</p> <p>1m</p>	<p>Award TWO marks for a circuit diagram drawn correctly with 2 bulbs, 2 cells and a switch [the components may be drawn in any order but the cells must be correctly orientated]:</p>  <p>If you are unable to award two marks, award ONE mark for a correctly drawn circuit which is missing one component or for a circuit which contains the correct components but there is one mistake in either the symbols used or how they have been connected.</p>	<p>TWO marks may be awarded for a non-rectilinear circuit or a circuit containing an obsolete symbol for a bulb:</p>  <p>Do not give full credit for a response that includes incorrect science:</p> <ul style="list-style-type: none"> circuits containing symbols not given or gaps between components of more than 2 mm circuits with extra/fewer components terminals on the cells facing each other circuits with incorrectly drawn components, eg: 	
Examples of learner responses				
	Entirely correct (2 marks)	Partly correct (1 mark)	Entirely incorrect (0 marks)	
				

Supplementary Material 5: Logistic regression with random effects - model fitting and checking

Model Fitting and Selection

Our final model was selected based upon Akaike's Information Criterion (AIC) (see e.g. Bolker *et al.*, 2009). Starting with the full model as described in the main paper, we considered dropping both fixed and random effects.

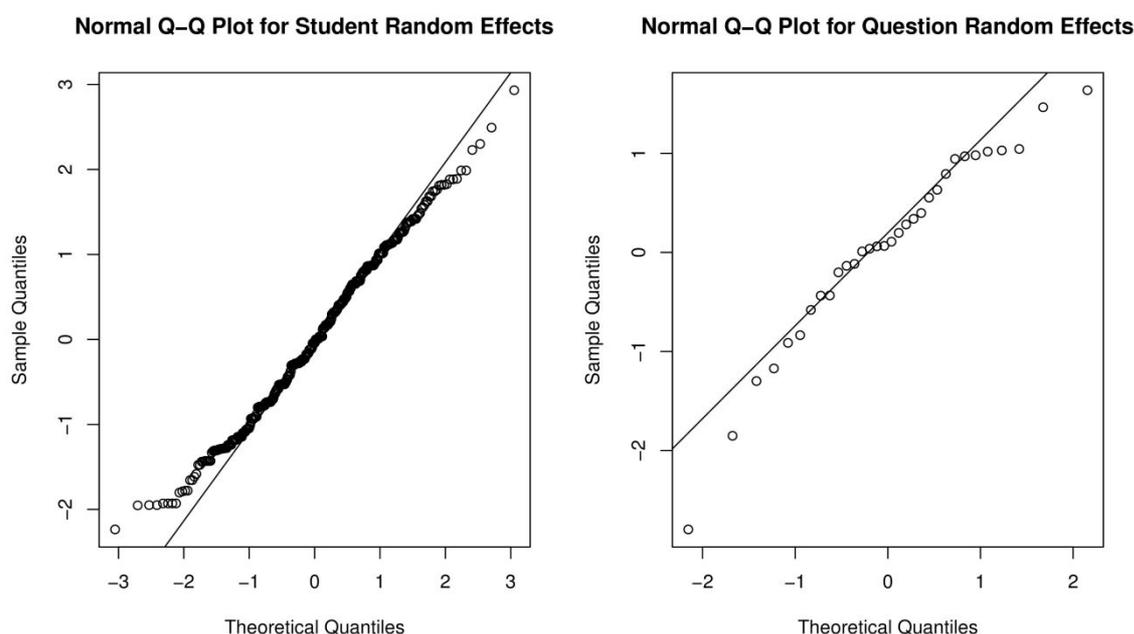
We also considered coarsening the English language proficiency class to both:

1. Binary category according to whether English was perceived to be their first language or not (i.e. ENS or ELL).
2. Binary category splitting learners into *proficient* English speakers consisting of native English speakers or those who self-reported that they spoke English "very well" (i.e. ENS or ELL level 1 learners); and *not proficient* speakers consisting those who were not native English and self-reported that they spoke English "OK" or "Not Very Well" (i.e. ELL level 2 learners)

The three level proficiency-based classification (i.e. ENS, ELL level 1 and ELL level 2 as discussed in the main paper) provided the lowest AIC suggesting that the model fit was significantly improved with the finer scaled proficiency classification and hence that the probabilities of responding correctly, and the effect of the question traits, were different between the three proficiency classes.

Model Checking

To check the random effects' normality assumptions, we provide below Q-Q plots of the random effects relating to the question number and student. The fit seems reasonable although there is some suggestion that the student random effects are somewhat light tailed and the question random effects left skewed.



Model Output

We provide a summary of the output of our final logistic regression model. The estimates provided are compared to the baseline of a native English speaker, in year 3, at a high ELL density school answering a "one beaker" *difficulty* question with a scientific fact (S) *focus* that requires *active* language production.

A positive estimate for a particular variable implies that changing from the baseline to this characteristic will increase the probability/odds of a correct (or partly correct) solution. A negative

estimate that it will decrease the probability/odds of a correct (or partly correct) solution. Note however that interactions terms will also need to be taken into account when comparing two hypothetical scenarios.

As explained in the main paper we fitted models considering both entirely correct and partly correct as our response variables. Little difference is seen in the results or resultant conclusions.

Using Entirely Correct as the response variable

Final Model Fitted

EntirelyCorrect ~ Year + Density + Difficulty + ProfClass + QuActive +
QuFocus + ProfClass * QuActive + ProfClass * QuFocus + (1 | StudID) + (1 | QuestionID)

where e.g. (1 | StudID) corresponds to a student random effect.

Fixed effects model terms – output for estimates β from R:

The baseline (intercept) corresponds to a native English, Year 3 student at a high ELL density school answering a one beaker, scientific fact question that requires active language creation.

Characteristic	Estimate	Std. Error	z value	Pr(> z)	Signif.
(Intercept)	-1.82870	0.40423	-4.524	6.07e-06	***
Year 4	0.17331	0.19636	0.883	0.377448	
Year 5	0.23419	0.20765	1.128	0.259413	
Year 6	1.06442	0.20460	5.202	1.97e-07	***
Low ELL Density	0.96611	0.23813	4.057	4.97e-05	***
Medium ELL Density	0.90361	0.15102	5.983	2.19e-09	***
2 Beaker Difficulty	0.07975	0.43434	0.184	0.854318	
3 Beaker Difficulty	-1.63007	0.50307	-3.24	0.001194	**
ELL Level 1	-0.05518	0.17183	-0.321	0.748083	
ELL Level 2	-1.05434	0.27002	-3.905	9.44e-05	***
QuActiveNo (No Active Language)	0.34501	0.39519	0.873	0.382647	
QuFocusR	1.86421	0.77809	2.396	0.016580	*
QuFocusSFV	-0.45302	0.57564	-0.787	0.431287	
QuFocusV	0.65620	0.70470	0.931	0.351758	
Interactions of Language with Active Language Creation					
ELL Level 1: QuActiveNo	0.00992	0.14509	0.068	0.945460	
ELL Level 2: QuActiveNo	0.73090	0.25609	2.854	0.004317	**
Interactions of Language with Question Focus					
ELL Level 1: QuFocusR	-0.27391	0.26650	-1.028	0.304039	
ELL Level 2: QuFocusR	0.11518	0.40492	0.284	0.776055	
ELL Level 1: QuFocusSFV	-0.05839	0.23111	-0.253	0.800525	
ELL Level 2: QuFocusSFV	0.52723	0.42088	1.253	0.210323	
ELL Level 1: QuFocusV	-0.87246	0.24842	-3.512	0.000445	***
ELL Level 2: QuFocusV	-1.33252	0.43626	-3.054	0.002255	**

Significance Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Using Partly Correct as the response variable

Final Model Fitted

PartCorrect ~ Year + Density + Difficulty + ProfClass + QuActive +
QuFocus + ProfClass * QuActive + ProfClass * QuFocus + (1 | StudID) + (1 | QuestionID)

Fixed effects model terms – output for estimates β from R:

SUPPLEMENTARY MATERIAL

The baseline (intercept) corresponds to a native English, Year 3 student at a high ELL density school answering a one beaker, scientific fact question that requires active language creation.

Characteristic	Estimate	Std. Error	z value	Pr(> z)	Signif.
(Intercept)	-1.22866	0.32036	-3.835	0.000125	***
Year 4	0.18095	0.18896	0.958	0.338266	
Year 5	0.29403	0.19941	1.475	0.140344	
Year 6	1.05061	0.19541	5.376	7.60e-08	***
Low ELL Density	1.00125	0.22999	4.353	1.34e-05	***
Medium ELL Density	0.98790	0.14604	6.764	1.34e-11	***
2 Beaker Difficulty	-0.15129	0.32215	-0.470	0.638631	
3 Beaker Difficulty	-2.00718	0.37854	-5.302	1.14e-07	***
ELL Level 1	-0.11060	0.16488	-0.671	0.502339	
ELL Level 2	-1.02894	0.24946	-4.125	3.71e-05	***
QuActiveNo (No Active Language)	0.42011	0.29508	1.424	0.154527	
QuFocusR	1.26794	0.58146	2.181	0.029212	*
QuFocusSFV	-0.99075	0.43255	-2.290	0.021995	*
QuFocusV	0.29246	0.52827	0.554	0.57984	
Interactions of Language with Active Language Creation					
ELL Level 1: QuActiveNo	0.04617	0.13774	0.335	0.737482	
ELL Level 2: QuActiveNo	0.62874	0.23778	2.644	0.008189	**
Interactions of Language with Question Focus					
ELL Level 1: QuFocusR	-0.20616	0.26313	-0.783	0.433340	
ELL Level 2: QuFocusR	0.15440	0.39848	0.387	0.698399	
ELL Level 1: QuFocusSFV	0.01424	0.22821	0.062	0.950243	
ELL Level 2: QuFocusSFV	0.58480	0.41299	1.416	0.156776	
ELL Level 1: QuFocusV	-0.76292	0.24394	-3.127	0.001763	**
ELL Level 2: QuFocusV	-1.31570	0.42482	-3.097	0.001954	**

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1