English language and literacy development of home and international students at a UK university

Justyna MacKiewicz

PhD

University of York

Education

September

2022

Abstract

Language and literacy skills are vitally important for academic success. In international contexts, an understanding of how these skills differ between the student populations is needed to put appropriate policies and support in place.

Three groups of first year undergraduate students at a UK university were recruited—English native speakers (ENS, N=59), first language (L1) speakers of one of the European languages (N=60), and L1 Chinese speakers (N=58). Their language and literacy skills (knowledge of vocabulary, grammar, reading comprehension, writing, and phonological skills) were compared at the start of Year one, while controlling for general cognitive abilities. One-way ANOVAs revealed significant group differences on all linguistic measures. Planned contrasts confirmed a large gap between the ENS and English foreign language (EFL) students, but also that this difference was primarily driven by Chinese students. Students with European L1s not only demonstrated stronger English skills than Chinese students, but also performed indistinguishably from British students on several tasks.

The same task battery was administered again after one year. T-tests and mixed-design ANOVAs were performed on participants tested at both time points. This sample included 48 ENS, 50 EFL with European L1s, and 47 EFL with Chinese L1s. The findings confirmed large and significant differences between the EFL with Chinese L1s and the ENS. The group of Chinese L1 students did not close the initial gap with the other EFL group as well. The group of EFL with European L1s, on the other hand, improved to a greater extent, and managed to close the gap with the ENS in almost all remaining measures over the span of one year.

These results show that the large differences in language and literacy skills observed previously are not inevitable. They also challenge the received view according to which EFL students' language develops rapidly and effortlessly.

Table of Contents

List of tables

List of figures

Acknowledgements

I would like to express sincere gratitude to my supervisor, Prof. Danijela Trenkic, thank you for your guidance, support, and thorough feedback on my work. I wanted to thank Dr Claudine Bowyer-Crane, Dr Sally Hancock, and Dr David O'Reilly for their continuous help and expertise. Thank you to Dr Sophie Thompson-Lee and Louise Shepperd for your help with data analysis and for being such wonderful friends throughout the whole PhD journey. My heartfelt thank you goes to my beloved sister Róża, my best friend and companion during lockdowns, the difficult times during which most of this thesis was created. Finally, I wanted to say massive thank you to all the wonderful participants who contributed their time and commitment to this study.

This thesis is dedicated to my mum, Alicja.

Author's declaration

I declare that this thesis, including all data presented in it, is original work and that I am the sole author. This work has not previously been presented for an award at this, or any other university. All data collection and analysis were carried out by the author, Justyna Mackiewicz, and all sources are acknowledged as references.

This work was supported by the PhD studentship funded by the Department of Education at the University of York.

Chapter 1: Introduction

1.1 Research context and study aims

The landscape of UK higher education (HE) has changed considerably over the past two decades due to internationalisation defined in terms of student mobility. In the past decade alone, the number of international students has grown by over 30%, from over 400,000 international students enrolled in UK HE in the academic year 2010–2011 to over 600,000 in 2021–2022. During the same period, the total size of student population stayed intact. There were 2.5 million students studying for a HE qualification or for HE credits at UK HE in 2010–2011, and this number stayed at the same level up to 2019–2020 academic year (HESA, 2023). This means that the proportion of international students in relation to the whole student body has increased. In other words, the ratio of international to home students in the UK universities has grown considerably over the past decade. This has led to an increase in linguistic diversity in UK universities, as the great majority of international students pursue their degrees with English as a foreign language (EFL). The proficiency in the language of instruction is vital for academic success because EFL students need to comprehend lectures, take an active part in seminars, take examinations, and produce written work in English. Therefore, they need to reach a certain level of proficiency to avoid language-related problems while pursuing their studies. The level of English they start with is also crucial in terms of the ultimate academic achievement (Trenkic & Warmington, 2019). However, little is known about the differences between international and British home students in their language and literacy skills at the start of university, how these skills develop, and how they impact academic attainment (Daller, Kuiken et al., 2021).

The proficiency in the language of instruction is crucial for academic achievement, as research shows that better-developed English at the start of university usually leads to a higher level of academic success in those for whom English is a foreign language (Bellingham, 1995; Cho & Bridgeman, 2012; Woodrow, 2006). In addition, knowledge of English and literacy skills can predict academic attainment in this student population. This includes knowledge of vocabulary (Daller & Phelan, 2013; Daller & Xue, 2009; Qian, 2002; Roche & Harrington, 2013), knowledge of grammar (Alderson, 2000), and skills in reading

and writing (Harrington & Roche, 2014; Li et al., 2010; Trenkic & Warmington, 2019). More recently, it has been demonstrated that phonological skills that underpin language and literacy are also an important factor in literacy development and ultimate attainment in English L2 adult populations (Schmidtke & Moro, 2020; Trenkic & Warmington, 2019). A limited mastery of English, on the other hand, diminishes international students' opportunities to learn and makes assessment difficult (Daller & Xue, 2009; Makepeace & Baxter, 1990; Paton, 2007). It is therefore important to investigate and compare the knowledge of English, literacy, and the skills that underpin them in students who pursue their studies in their first language and those for whom English is a foreign language.

At the same time, research shows that the academic attainment of international students is lower than that of home students in English-speaking countries, but the reasons for this are not well understood (Crawford & Wang, 2015; Iannelli & Huang, 2014; Morrison et al., 2005). Second language skills have been called into question as a possible explanatory factor on the grounds that international students meet the strict language requirements set by universities. However, the latest evidence points to large and significant differences in language and literacy skills between international and home students at the start of university and shows that language predicts academic outcome to a great extent until a certain level of proficiency is reached. Trenkic & Warmington (2019) compared English language and literacy skills in British home students and international students from China on a number of language measures. The findings showed that Chinese L1 students had substantially smaller vocabulary, read at half the speed, understood less of what they read, and were less able to summarise what they had read in writing. Crucially, the initial differences persisted over the course of the first year at university, and individual differences in language skills were predictive of academic success of international students but not of ENS students. Reaching the minimum entry requirements set by the university did not allow Chinese L1 students to perform to the level of their ability, which indicates that many international students pursue university education at a systematic disadvantage.

Trenkic and Warmington's study (2019) suggest that Chinese L1 students may be at a striking disadvantage regarding the language skills essential for academic success. However, it is not known to what extent these findings can be generalised to other international

student populations. The main goal of this study is therefore to compare language skills in home and international students with the international student populations different from each other in as many aspects as possible (including linguistic and cultural aspects, and geographical distance from the UK), to validate the findings from Trenkic & Warmington (2019). The group selected for comparison consisted of international students from Europe, as they are different from Chinese L1 students in all these dimensions making them a fair sample for a validation study.

In response to limited evidence on the differences in language and literacy skills between international and home students, and building on the previous findings, this study involves an approximate replication of Trenkic & Warmington (2019) with the aim to investigate language and literacy skills in undergraduate university students at a UK university. The following research questions (RQs) are addressed:

RQ1. How much do knowledge of the English language and literacy skills differ at the start of the first year at university between:

a) Home students who speak English as their first language and international students who speak English as a foreign language (EFL)?
b) Those speaking European L1s and Chinese L1s?

RQ2. How much do knowledge of English and literacy skills change in the first year at university in home students and in international students speaking Chinese and European L1s? Do international students close the gap on any measures?

1.2 Importance of the study

It is important to understand the linguistic differences between those who study in their native language and those who do so in a foreign language because all students work and are assessed against the same criteria while in an English-medium university. It is crucial to find out if they are all on a level playing field when competing against strict requirements set by universities. This in turn will inform those international students whose future educational and professional goals hinge on their degree classification level. A good degree classification and strong English skills are crucial when pursuing further studies at master's

and PhD levels in English-medium universities and entering a competitive job market in English-speaking countries. What is more, lack of strong skills in English may jeopardise these students' professional opportunities by causing them to be perceived as less intelligent, competent, and professional (Fuertes et al., 2012; Gluszek & Dovidio, 2010). It is therefore important to compare international and home students at the start of their degrees and to track language development, linguistic gains, and their importance for academic achievement. The results of this study will inform international students pursuing these goals, namely, a good degree classification and substantial improvement in English. The findings are also important for English-medium HE institutions, as they will help in identifying those cohorts that might be at the greatest linguistic risk and in developing appropriate language provision to mitigate these issues.

1.3 Outline of the thesis

This chapter has outlined the research context, aims, and importance of this study. Chapter 2 presents the background literature leading to the identification of research gaps. It consists of four main parts. The first part discusses the importance of international students in UK HE, the level of their academic attainment at English-medium universities, and the level of English they arrive with. The second part demonstrates the importance of language in academic attainment and points to some specific types of knowledge and skills that are important at university. The third part narrows the scope to international students from China, discussing their academic attainment and linguistic issues. Finally, the fourth part highlights the linguistic differences between English and Chinese that may be influential in terms of learning while at university. Chapter 3 then presents methodology and methods for the original design used in the first of two waves of data collection. It describes the participants, the task battery administered at the first time point, and statistical analyses for data obtained at both time points. The results based on the first wave of data and pertaining to the first research question are presented in Chapter 4. Chapter 5 describes the revised procedure used in data collection at the second time point and administered in response to fieldwork limitations caused by the Covid-19 pandemic. This chapter also presents the results based on the second wave of data that pertain to the second research

question. A discussion of all the findings is provided in Chapter 6, followed by a conclusion in Chapter 7.

Chapter 2: Literature review

2.1 Chapter outline

This chapter starts with a short section providing definitions of the most important terms used throughout this work. It is followed by a brief presentation of the demographics and importance of international students in UK universities (section 2.3) to set the scene for this study. It shows that their academic attainment at university is lower when compared to that of native speakers of English and that they begin university with a considerable linguistic disadvantage that may be responsible for these attainment differences. Some evidence supporting the link between language and academic achievement will be provided, based on comparative studies of attainment differences. This will be followed by a critical evaluation of two counterarguments undermining the relationship between proficiency in English and academic achievement. The first one rejects English proficiency as a possible factor on the grounds that international university students meet the language requirements upon entering universities, and the second one rejects the idea of attainment differences altogether.

Section 2.4 reviews the literature demonstrating the importance of English in academic achievement. It starts with the predictive validity of proficiency tests such as IELTS and TOEFL. Another strand of research on the importance of English in academic achievement is based on studies showing the importance of vocabulary, grammar, and reading and writing skills in English. This section also briefly describes some cognitive abilities that are important in learning and language development. The next section (2.5) narrows down the scope, turning to studies on the challenges faced by international students speaking Chinese as their L1, whose linguistic difficulties have been particularly well documented. It will show that, on the whole, they struggle linguistically the most and that their academic attainment is lower when compared to that of other international students. Section 2.6 introduces the notions of language family and linguistic transfer, and outlines the main differences between Chinese and English. It shows that Chinese is a language very different from English and that these cross-linguistic differences may place an additional learning burden on this group of students while at university.

2.2. Definitions

The term *international students* refers to students pursuing their studies outside of their home country and is used to distinguish them from local, home students. According to UNESCO (2019), '[a]n internationally mobile student is an individual who has physically crossed an international border between two countries with the objective to participate in educational activities in a destination country, where the destination country is different from his or her country of origin.' This term, however, can refer to different student populations as it depends on the study context, and consequently, it is used in various ways in the literature. In the UK, home and international students are distinguished based on their domicile, with international students being those studying in the UK but domiciled elsewhere. On this basis, students of foreign nationality making their home in the UK are treated as home rather than international students. In the UK, international students are further subdivided into EU and non-EU students. The EU students are those domiciled in one of the European Economic Area (EEA) countries, and all other international students are referred to as non-UK students.

With respect to the first language spoken, university students in English-medium HE consist of two main groups: English native speakers (ENS) and English foreign language (EFL) students. The ENS and EFL students are sometimes referred to as home and international students respectively; however, not all home students are English native speakers (e.g., UK-domiciled Italian nationals), and not all international students are English L2 learners (e.g., Australian students in the UK). Therefore, the following nomenclature is used in this study: *international students* is an umbrella term referring to all students pursuing their studies outside of their home country irrespective of their domicile and nationality. *Home students* or *domestic students* are terms used to refer to non-international students. *ENS home students* or *British home students* are home students who claim English to be their first and dominant language. *English foreign language students* are those home and international students who speak a first language other than English and who have learned English mostly as a school subject through formal instruction. The EFL students are further divided into those speaking a European language as their first language (*EFL students speaking European L1s*) and those speaking Chinese as their first language (*EFL students speaking Chinese L1s*).

There are two more concepts mentioned throughout this thesis that can carry nuance in their meaning and for this reason, their intended meaning is explained here. The first one is the *threshold* in language proficiency. It refers specifically to the level of proficiency in English that students would ideally meet, and once this is achieved, the importance of English language proficiency can be suppressed by other factors (Graham, 1987). Also, higher scores in the language entry examinations are associated with a higher level of achievement until this threshold in language proficiency is reached. The second term, *academic successes*, when used in the context of EFL university students, means obtaining grades and degree classification equivalent to those that would be obtained by students if they pursued their studies in their first language. In other words, a success would mean obtaining grades students are capable of in their first language.

2.3 International students in the UK, their language and academic achievement

2.3.1 Importance of international students in UK HE

The number of internationally mobile students tripled from 2000 to 2019, from two to six million (UNESCO, 2022), and a great majority of them study in their second language. International students select traditionally English-speaking countries such as the USA, the UK, and Australia (Manning, 2018), but also other non-English speaking countries that offer English-medium programmes, making English the dominant language of HE (Wächter & Maiworm, 2014). The United Kingdom (UK) is the second most popular destination for international students in the world. As of 2017 it hosted 8.1% of international students, just below the USA, which received 17.7% of the total international student number (Chervenkova, 2021). In the academic year 2019–2020 there were around 540,000 international students enrolled at UK HE institutions. This figure accounted for 27.5% of the total student population in the UK, 15.7% of all undergraduate students and 40.3% of all postgraduate students (Chervenkova, 2021).

The total population of international students in the UK in 2019–2020 comprised 142,990 EU students (of whom 70.5% were undergraduates) and 395,630 non-EU students (of whom 47.3% were undergraduates). The greatest proportion of international students arrive from

China and from EU countries, with Italy being the top sending country within the EU. The number of students from China has been growing at the fastest rate, with a 100% increase in the 5-year period between 2014–2015 and 2019–2020. In 2019–2020 there were 104,240 Chinese students at UK HE institutions, making up 28.9% of all international students (Chervenkova, 2021). International students comprise a heterogeneous sample in terms of the subject studied. According to the latest statistics, the most popular disciplines among these students are business and management, and the majority select subjects within the humanities and social sciences (Chervenkova, 2021). The total number of international enrolments in the UK has grown considerably over the past years, with an increase of 50,000 between 2018–2019 and 2019–2020 alone, the largest increase in the decade. UK government reports estimate that there will be 600,000 international enrolments every year by 2030 (International Education Strategy, 2021).

The rapid increase of international enrolments brings financial benefits to the receiving HE institutions. In 2018–2019 international students contributed £15 bn to UK institutions in tuition fees alone (Chervenkova, 2021). The universities enjoy the benefits of international students' fees, especially from full fee-paying non-EU students. In 2018–2019 those students spent £12.5 billion on fees alone, and it is argued that these students' fees provide a vital cross-subsidy to the teaching of home students and to research, without which it would be much harder for British universities to remain competitive in the world market. A study by Olive (2017) analysed cross-subsidisation and found that on average, each non-EU student subsidises research by over £8,000 over the duration of their study. This income would not only be used to support the funding of research; it can be used to support investment in university facilities and the sustainability of certain courses and modules for home students that would not be financially viable if not for the presence of international students. It is claimed that many existing courses would be at risk without them.

International students also contribute to the UK economy through their spending on living expenses, which helps to sustain economic activity and employment. The sectors that benefit the most from the presence of international students are real estate, transport, and retail. Additional economic activity is generated by the visiting friends and family of students through their spending on transport, hotels, hospitality, and cultural, recreational, and

sports attractions. In 2018–2019 international students generated £13 billion of income (excluding tuition fees), and the income from students' visitors and family amounted to £0.7 billion. Combining the direct and indirect benefits of tuition fees, spending on living expenses, and visitors' spending, the total benefit to the UK economy in a single year was approximately £95,000 per EU-domiciled student and £110,000 for a non-EU student (Chervenkova, 2021). The total costs associated with teaching grants, student support, and other public costs amounted to £2.9 billion, which shows that the benefits associated with the presence of international students outweighed the costs considerably (HEPI, 2021).

Aside from the financial benefits brought to HE institutions and the wider economy, international students also bring pedagogical benefits for the local students, who can experience other languages and cultures at their doorstep. They can develop international networks and gain intercultural understanding and communication skills vital in their future workplaces, which also become more culturally diverse. In one of the studies carried out in the UK, it was found that the diverse international context prepares all students to develop competences and tolerance which are likely to widen their employment and career opportunities (Montgomery, 2009). This qualitative study was based on data obtained in mixed-nationality focus groups, with 70 participants (33 international and 37 British students) from three programmes (business, design, and engineering) interviewed across 12 focus groups with a semi-structured interview. The participants reflected on the impact of working in international learning environments and specifically of mixed-nationality team working. The study found that international experience helped British students to gain different perspectives on the subject studied and increased their awareness of doing things in many different ways. The study concluded that all students developed awareness of the complexity of culture and noticed the diversity existing within a single culture. The participants agreed that the international experience at university prepared them for work in international contexts in the future.

International students also enrich the local learning environment through the exchange of ideas, as shown in a study by Luo & Jamieson-Drake (2013). Their evidence was based on alumni survey data from three cohorts of US citizens who had graduated from four US universities roughly 5, 10, and 20 years earlier. The survey, administered in 2005, included

questions on the level of involvement in college life, interactions with international students, beliefs and values during college life, and different areas of possible development while at university. The alumni were grouped into three categories according to the level of international interaction, varying from none or little, through some, to substantial. According to the findings, the US students who interacted with international students the most were those who questioned their own beliefs and values the most when compared to those at lower levels of interaction. Those students challenged their political beliefs and their beliefs about other religions, races, ethnicities, and sexual orientations more than their less interactive peers. They also indicated significantly higher levels of skill development, such as in reading and speaking another language, acquiring new skills and knowledge independently, and formulating creative and original ideas.

Internationalisation defined in terms of student mobility also affects local students' cross-cultural awareness. Geelhoed et al. (2003) investigated the benefits that home university students gained in a programme that paired a host US student with an incoming international student and that helped international students in their adjustment to the new country. This qualitative study explored the views, experiences, and feelings of 16 host students in focus groups. According to the findings, the programme had some cognitive influence on the host students as they gained new cultural perspectives, developed empathy towards the international students they were paired with and even managed to influence their family and friends' attitudes towards international students. They also became more competent with intercultural interactions, more aware of their biases and stereotypes, and more critical of their cultural assumptions.

The international students in English-medium HE also benefit from the experience of studying for a degree abroad. They have the unique chance to experience the host country and its culture, and improve their employment and career outcomes in the long term. However, one of the most desirable benefits is improvement in second language skills (Evans & Morrison, 2011). According to Coleman (2004), improvement in linguistic skills is the single most expected gain and for this reason '[s]tudy abroad is often integrated into degrees in modern languages... in the belief that extended immersion in a society where the target language is used every day will enhance the learner's proficiency, especially oral-aural

skills and less formal registers' (p. 582). This belief in rapid language improvement has very little empirical support. Little is known about the amount of language development gained while studying for a degree in an English-speaking country and the findings are not conclusive (Storch & Hill, 2008; Trenkic & Warmington, 2019).

Besides improvement in second language skills, another highly desirable outcome of studying abroad is an internationally obtained degree which brings advantages in income, status, and possible influence (Mellors-Bourne et al., 2013). However, research shows that international students obtain fewer good degrees when compared to the local, home students (this will be discussed in more detail in section 2.3.2), and the reasons for this are unclear. The two most desired benefits, namely improvement in second language proficiency and a good degree classification, are closely connected because language is at the heart of how much international students can achieve academically (this will be discussed in more depth in section 2.4).

It can be concluded that the student makeup in UK HE has changed considerably over the past decade. The traditional landscape of British students who are homogeneous in terms of their first language has been enriched by international students who bring benefits for the receiving institutions in the form of tuition fees, broad economic benefits through international students' spending, and pedagogical benefits to the local home students. It is therefore important to preserve the existing benefits through sustainable international enrolment. This can be achieved if the benefits for international students are fully recognised and the potential obstacles faced by international students in pursuing their goals are attended to by the receiving institutions and the students themselves. For these reasons, more research is needed into international students' language development and identifying potential problem areas, which can in turn feed into pedagogical and policy-related decisions. This is aimed at developing a more level playing field with regard to academic attainment in the sense of equality and opportunity. Since both aspects – namely, academic achievement in the university course and the level of English at the start of the course – are closely connected, they will be discussed next. The next two sections will discuss the attainment differences observed in home and international students (section 2.3.2) and the level of English with which international students begin their studies (section 2.3.3).

2.3.2 Academic achievement in home and international students

Good grades and consequently a good degree classification are amongst the most highly desirable benefits for all university students. In the UK, a bachelor's degree classification is based on a weighted average of the marks obtained in exams and other assessments from the second year at university onwards. Those who obtain an average mark of at least 70% in their exams obtain a first-class degree. The second-class degree is divided into upper second class (average marks of 60–69%) and lower second class (average marks of 50–59%). The third-class degree is at the lowest end and is obtained by those who score on average between 40% and 49%. First-class and upper second-class degrees put students in a good position for employment, graduate programmes, and postgraduate study. Despite the importance of academic achievement, comparative studies of international and home students show consistently that international students obtain lower exam marks and fewer good degrees. This has been found at universities in English-speaking countries around the world, including Canada (Berman & Cheng, 2001; Fox, 2005; Grayson, 1997; Grayson, 2008a, 2008b; Grayson & Stove 2005; He & Banham, 2009; Morris & Cobb, 2004), Australia (Murray, 2010; Paton, 2007), New Zealand (Read, 2008), and the UK. Research from the UK shows that international students obtain fewer good degrees when compared to home students (Crawford & Wang, 2015; Iannelli & Huang, 2014; Morrison et al., 2005), perform less well in the end-of-year examinations (Brackenbury, 2019; De Vita, 2002; Li et al., 2010), and are more likely to fail their degree programme (Makepeace & Baxter, 1990). Many factors have been proposed to explain this phenomenon but none of them has explained the variance in marks.

For example, Morrison et al. (2005) showed that undergraduate international students in the UK obtain fewer good degrees when compared to home students. The study used data collected centrally by the Higher Education Statistics Agency (HESA) to investigate whether there were differences in performance between home and international students in terms of the degree classification obtained and the factors contributing to the potential differences in academic attainment. The data came from 15 universities for undergraduate students in the years 1995 and 2000. The results showed that the overseas-domiciled students obtained fewer good degrees (i.e., first-class or upper-second class ones) when

compared to home students. None of the demographic factors investigated explained the variance in marks, but as the great majority of international students pursue their studies with English as their second language, it is likely that their level of skill in English could have impacted their attainment. This claim finds some support in the regional variation in findings, as the students from North and South America, non-EU Europe, and Australasia, regions inhabited largely by native speakers of English, did not vary significantly in performance from home students in the UK. However, those domiciled in the EU, Asia (specifically in China, where most international students come from), Africa, and the Middle East performed less well than the home students.

A further investigation of factors responsible for the attainment differences between international students from China and home students in the UK was undertaken by Crawford & Wang (2015). This study was based on a sample of 112 undergraduate students enrolled in a single HE institution. The sample consisted of Chinese L1 students (*N* = 52) and home students (*N* = 60) enrolled in three- or four-year programmes in Accounting and Finance in the academic years 2006–2007 and 2007–2008. The study investigated the importance of gender, prior academic achievement, prior academic qualifications, degree programme, and enrolment year on yearly marks and the final degree classification. The results show that even though the Chinese L1 students outperformed the UK students at the end of the first year at university in both enrolment years, their performance worsened in years two and three in relation to the home students. The difference in the final degree mark between the two groups in question was statistically significant, with 80% of the UK students and 43% of the Chinese L1 students graduating with a good degree.

A further finding was that the performance of the Chinese L1 students was not affected by any of the predicted factors. Instead, these factors proved to be important in predicting outcomes in the sample of UK students only. This suggests that there exists a different set of factors predicting academic achievement for each group. One of the main differences is that international students pursue their studies in their second language; given the importance of this, it may be the reason behind the outcome differences. Chinese L1 students were able to compete with the native speakers of English in the first year, but from the end of the second year their performance worsened in relation to the home students, and the gap widened further by the end of the final year at university. The reason

for this may be that international students reach the limit of their language resources from the second year onwards when the content of the subjects studied becomes more challenging, making it more difficult to keep up with the native speakers of English whose skills were constantly developing.

International students also seem to obtain lower examination marks when compared to home students, and the examination type was considered as one of the possible factors in a study by De Vita (2002). This investigation was based on first-year students pursuing degrees in business and management in the academic year 1999–2000. The sample consisted of home students (*N* = 195) and international students (*N* = 109) of 24 different nationalities. The data were based on the final marks from one module, comprising different types of assessment: group work involving poster presentation, an end-of-term multiple-choice test, and a closed-book final examination comprising a mixture of short answers and essay-type answers. The results showed that home students outperformed their international counterparts consistently in all three types of assessment: the multiple-choice test (67% vs 55%), the coursework assignment (59% vs 53%), and the final examination (62% vs 41%). The multiple-choice test and final exam were thought to pose the greatest demands because of their timed nature. The authors suggested that results for these methods of assessment were likely to be influenced by the cultural and linguistic characteristics of international students because of the slower processing of text written in students' L2 which could have extended the time needed to take these exams. This issue has been recognised by other researchers (Mendelson, 2002) who recommend longer time limits in exams for students speaking L2 English.

Differences in course marks were also found in a study investigating the effect of lecture capture on students' performance (Brackenbury, 2019). Lecture capture is an audio recording of a lecture that students can access after the lecture, and it serves as a resource in exam revisions. The usage of these recordings was operationalised twofold as the total number of views and the total number of minutes spent on listening. Data on lecture capture usage, nationality, gender, and disability status (including language-related disabilities) were collected from students in the second and third year of a three year BSc honours programme. The examination marks (on a scale between 0 and 100) came from

eight bioscience modules taught in the academic year 2017–2018 at the University of York. According to the results, use of lecture capture did not predict exam marks and those who used it more extensively did not perform better than those who used it to a lesser degree. The factors that predicted students' marks turned out to be nationality and disability status. The UK students (*N* = 763) performed significantly better than the non-UK students (*N* = 79), obtaining a mean mark of 61.2 (SEM= 0.50) and 57.2 (SEM=1.80) respectively. Students with no recorded disabilities performed significantly better than students with such disabilities. The author concluded that the disabled and non-UK students were disadvantaged, making further enquiry into their academic achievement necessary.

More evidence on attainment differences between international and home students comes from other English-speaking countries, corroborating the findings obtained in the UK. A similar pattern of performance was found in a study investigating the effect of academic experience on academic achievement in first-year undergraduate international and home students at four Canadian universities in the academic year 2003–2004 (Grayson, 2008a). The whole sample consisted of 1,415 students with the international subsample originating from China, Hong Kong, and Taiwan. Academic experience was defined as engagement in formal and informal campus activities, whereas the outcomes were measured via self-assessment, completed credits, and grade point average (GPA). The GPA is simply an average mark obtained by adding up grades and dividing them by the number of grades. It is one of the most common measures that quantifies academic achievement, and it will be cropping out on many occasions throughout this work, especially in section 2.4 discussing the importance of language in academic achievement. The study found that the international students performed less well than Canadian home students in both self-assessed and objectively measured outcomes. For the self-assessment outcomes, a significantly smaller proportion of international students stated that they were satisfied with their academic outcome (70% vs 75%) and believed that they had developed intellectually over a year (61% vs 74%) and increased their knowledge over a year (78% vs 88%). Furthermore, the home students completed significantly more credits and obtained significantly higher GPAs when compared to the international group.

The above review shows differences in the level of academic achievement between home and international students, but none of the investigated factors explained the variance. These findings therefore beg the question of why these differences exist. In addition, the factors investigated in the above-mentioned studies seem to predict the outcomes in the two groups in different ways. For example, in the study by Grayson (2008a), students' background and academic experience predicted academic achievement in home students to a greater extent than in international students. This was also true in a study by Crawford & Wang (2015), where the demographic factors investigated were predictive in terms of academic achievement only in the sample of home students, but not for the international students. This suggests that these two groups are somehow different. One of the most obvious factors that differentiates the two groups of students and that can hinder their studies are their skills in English. International students use their second language and, given the importance of language in learning at university, this may have an impact on their academic attainment.

When language is considered, it seems to influence academic attainment in international students (Berman & Cheng, 2001; Brooks & Adams, 2002; Fox, 2005; Makepeace & Baxter,1990; Trenkic & Warmington, 2019; Volet & Renshaw, 1996). In one of the earliest studies on attainment differences, Makepeace & Baxter (1990) investigated 153 failed overseas students from 15 polytechnics enrolled in the academic year 1987–1988 and examined factors associated with the students' failure. These factors included entry qualifications, language, numeracy, coursework, attendance, motivation, study habits, lack of ability, finance, immigration, medical factors, accommodation, and personal factors. Each of those factors was rated by the teaching staff in terms of their contribution to attainment differences and study success. Language proved to be the most important predictor of students' failure as it was mentioned as the factor affecting the outcome by 40% of the respondents.

It was also found that overseas students in their first year in the polytechnic sector did not perform as well as their UK counterparts. This conclusion was based on the numbers of home and overseas students who failed their courses, and it was true at all levels, from undergraduate to master's level courses. The authors concluded that '[l]anguage, in particular comprehension appears to be the most common explanation for academic failure'

and that 'failure is as likely to occur in students who have been in the UK for up to 5 years, as it is for new arrivals' (Makepeace & Baxter, 1990:38). The fact that language could have been responsible for some of the international students' failure is further supported by the fact that only 6% of the international students who failed their courses held some formal qualifications in English.

The importance of English in academic achievement ascertained by the teaching staff seems to be corroborated by the perceptions of international students themselves. Berman & Cheng (2001) investigated the academic performance of English L2 undergraduate and postgraduate students in relation to their native-English-speaking peers after the first semester at a Canadian university in the academic year 1998–1999. The data were obtained from a sample of non-native speakers of English (*N* = 113) and native speakers of English (*N* = 73) and were based on self-reported difficulties in reading, writing, speaking, and listening. The perceived levels of difficulty of the four language skills were then correlated with students' GPA obtained after the first semester. According to the results, there was no relationship between any of the language skills and the GPA in the native-English-speaking sample, in both undergraduate and graduate students. The relationship was found only in the group of English L2 graduate students. It was found that for all four language skills, the more difficult the skill was perceived to be, the lower the GPA tended to be. The academic performance of this postgraduate group was significantly correlated with 27 out of 40 language items in the questionnaire. In sum, language was found to be related to the grades obtained by the postgraduate English L2 students to a great extent.

Attainment differences were also found between home and English L2 students who had spent a considerable amount of time using English in school prior to the start of university. Fox (2005) investigated the relationship between the amount of time English L2 students spent in an English-medium high school and their success at university in Canada. The English L2 participants were enrolled in undergraduate courses in engineering and commerce and admitted to university based on the language residency requirement that exempts students from proving English qualifications upon completing 3 to 5 years in an English-medium high school. The sample comprised 265 students speaking around 40 different first languages, with 19% Mandarin L1s, 6% Tamil L1s, and 3.7% Farsi L1s.

The results showed that the language residency group obtained significantly lower GPAs at the end of the first year when compared to the means typically reported at this university from the previous 5 years. Their grades were also lower when compared to other English L2 students who had passed an English language entrance test and who were supported linguistically during the whole course of their studies. Only 14% of students in the language residency group completed their second year at university successfully. A follow-up qualitative analysis of interviews with 4 professors and 15 students from the same institution confirmed that limitations in English skills constitute one of the main obstacles to the successful completion of a degree, specifically difficulties with productive skills such as speaking and writing.

The relationship between productive skills in English and academic achievement was investigated in another study in an Australian university. Brooks & Adams (2002) investigated the effect of language on the examination results after the first semester among first-year business students. This study was carried out in response to some concerns over home students for whom English was an additional language and English L2 international students. The researchers administered a survey on familiarity with spoken English to investigate oral comprehension and spoken English. The participants (*N* = 285) were asked about the frequency of using English. Academic achievement was conceptualised as the mark obtained in one module (out of 100) and based on a presentation, a project, and two examinations. The results showed that home students used English at a significantly higher rate when compared to international students and they obtained significantly higher grades at the end of the semester. The authors suggested that the attainment differences may originate in an inadequate level of English. However, there may be other possible factors that are not accounted for in the study and these findings need to be treated with caution.

The above review shows that, when English language skills are taken into consideration, they tend to be related to academic outcome in international students (Berman & Cheng, 2001; Brooks & Adams, 2002; Fox, 2005; Makepeace & Baxter, 1990). This suggests that there exists a threshold in proficiency and language competence is a necessary but not sufficient condition for academic success. This has been noticed by other researchers: for

example, according to Cho & Bridgeman (2012), '[o]ne obvious example that attests to this is that not all native speakers are academically successful. By the same logic, there is no reason to expect that all NNES [non-native English speakers] students with high levels of English proficiency would necessarily have high GPAs' (p. 424).

Similarly, Elder et al. (2007) stated that '[if] language proficiency were all that mattered, then native speakers would be automatically assured of an easy passage through their academic courses, regardless of their level of disciplinary knowledge or of other attributes such as intelligence, initiative and effort which are rightly rewarded in the academic domain' (p. 53). This poses the question of what the necessary level of English is that allows international students to pursue HE studies alongside native speakers of English and that does not affect their studies in a negative way. The next section sheds some light on what can be the sufficient level of English skills that is necessary for international students at the beginning of university to be able to compete against native speakers of English. It is followed by an overview of studies on the linguistic abilities international students begin their studies with.

2.3.3 International students' proficiency in English at the start of university

Language and literacy skills, and their development and impact on academic attainment, have been studied extensively in the context of immigrant schoolchildren with English as an additional language who start their education with limited English proficiency and a significant gap dividing them from their native peers. There is strong evidence showing that these students cannot compete with their native English peers until they catch up with them linguistically. According to Hakuta et al. (2000) it takes immigrant children between 3 and 5 years to develop basic interpersonal communication skills (BICS). However, those skills are not sufficient in an academic learning context, where language puts a great demand on students. To succeed in an academic context, students need to develop cognitive academic linguistic proficiency (CALP). There is evidence from studies on bilingual immigrant children suggesting that it may take between 5 years (Cummins, 1981) and 'up to 10 years before students are fully proficient in English, i.e., are fully competitive in the academic uses of English with their age-equivalent, native English-speaking peers' (Hakuta et al., 2000:1).

Language and literacy development seems to be a long process also in adolescents who enter secondary school with limited proficiency in English. Some scarce evidence available for this age group showed that students in grades 9 and 10 with two years of immersion in an English-medium school, after entering these schools highly literate in their first languages, demonstrate significantly lower skills in English when compared to their peers (Pasquarella et al., 2012). The participants in this study consisted of native English speakers (*N* = 31) and non-native English speakers (*N* = 44) of 17 different nationalities. They had lived in Canada and attended schools there for over two years at the time of taking part in the study. The two groups were compared in their decoding skills, knowledge of vocabulary, and reading comprehension. The results showed that English L2 adolescents' scores were up to 3 standard deviations below the range of the scores obtained by English L1s. Their skills in reading were also influenced differently. It was the knowledge of vocabulary that was predictive of reading comprehension in native speakers of English. Reading comprehension in English learners was predicted by both decoding skills and vocabulary. These findings show that the difference between English L1 and L2 students is still significant after 2 years at an English-medium school, and that L2 students' reading skills were still developing at this stage.

The findings from the context of children and adolescents with English as an additional language cannot be generalised to the university student population because they constitute qualitatively different samples. However, they are still relevant here as theoretical underpinning on language and literacy development. Notable differences between schoolchildren and university students are that university students are cognitively mature at the start of their studies, they have well-developed L1 skills that facilitate second language learning, and they have mastered their second language to the level necessary to commence university studies. What is more, they demonstrate qualifications in English required by universities. Despite all these characteristics, the level of language competence with which university students commence their studies is still significantly lower when compared to native speakers of English. However, there is a lack of longitudinal studies that show how language develops in this context over time, whether the initial gaps close during university studies, and how these differences translate in terms of ultimate academic achievement.

Research on linguistic differences between native speakers of English and EFL students at the start of university shows that there exist large and significant differences in vocabulary knowledge (Elder et al., 2007; Morris & Cobb, 2004). For example, Morris & Cobb (2004) evaluated the role of vocabulary profiling at entry to a university course in identifying students who may be at risk of failing their degree due to linguistic problems. The data used in analysis consisted of entrance exams for a teacher training course in Canada. These exams were completed by native speakers of English and English L2 learners (*N* = 122) representing 14 different nationalities (the majority being French, Greek, or Italian). The analysis of vocabulary in these assignments showed that native speakers of English relied less on the first thousand of the most frequent words in English and on function words. They also used more academic vocabulary when compared to English learners. All these aspects of writing indicated that native speakers' work was of significantly better quality when compared to that of their English L2 counterparts, and the vocabulary profiles predicted participants' performance in the course they were enrolled on. These results show that the gap in vocabulary knowledge that exists from the start has a real impact on students' performance. It is important to note that all EFL students met the linguistic entry language requirement, with many of them being rated as having near-native proficiency.

In another study by Elder et al. (2007), the authors assessed the language of a diverse population at entry to a New Zealand university and its importance in academic achievement. The data came from a two-part diagnostic test, with a receptive vocabulary test designed to filter linguistically able students out from further diagnosis. Students taking part in the second step were further assessed on academic reading, listening, and writing and assigned a score on a scale between 1 and 6. A score of 4 or 5 indicated a student who was at risk, and a score of 6 meant that the student's language skills were inadequate to pursue further studies. The results were obtained from native speakers of English (*N* = 1,257) and English learners (*N* = 1,785) from a range of subjects. The study found that out of the students undertaking the second part of the assessment, over half of the English learners were assigned scores of 4, 5, or 6, meaning that they were at-risk students. A relatively small proportion of native speakers of English obtained low scores in this assessment. A further analysis demonstrated that students obtaining scores of 4 or 5 were

three times more likely to fail their course. This study shows that international students were are at greater risk of failing their university course than the native speakers of English.

Large and significant differences were found between ENS and EFL students on a range of language tasks at the start of university in Canada (Devos, 2019). The sample of ENS consisted of 319 students, and the English learners (*N* = 118) represented 22 different first languages, including Mandarin (25%), Korean (22%), Vietnamese (13%), and Cantonese (7%). All the participants were further divided into three groups: those exposed to English between birth and the age of 5, early learners (6–17 years old), and late learners (17+ years old). Participants' skills in reading, writing, and vocabulary knowledge were assessed at the start of the first semester. The results showed that performance in these three tasks was predicted by the length of using English, with students who had acquired English since birth performing significantly better than the other two groups of students in all three tasks: reading, writing, and vocabulary. No significant differences were found in vocabulary and reading between the early and late groups, but significant differences were found between all three groups in the writing task. The findings suggest that vocabulary, reading, and writing take a long time to develop, as even those who started learning English at the age of 6 were significantly different from native speakers of English at the point of entering university. In addition, those residing in Canada for up to 10 years obtained significantly lower scores across all tasks when compared to those who had resided in Canada for more than 11 years.

There exists only one study so far that not only compared the initial differences in language and literacy skills in international and home students, but also compared the skills that underpin them, tracked the development of all these skills over the period of one academic year, and showed their relationship with academic achievement. The study by Trenkic and Warmington (2019) consisted of international students from China (*N* = 63) and British home students (*N* = 64). They were compared on several language and literacy skills, and also on skills that underpin literacy development such as phonological skills and spelling. The findings corroborated large and significant differences between home and international students in all measures at both time points. All the measures explained 51% of the variance in marks of the Chinese L1 sample and only 10% in the British student sample. The phonological skills (unexplored in earlier studies) correlated significantly with academic

achievement at the end of the first year in the Chinese L1 sample and they were measured in a task that required participants to delete a specified phoneme from a word to create a new word. The spelling skills predicted a portion of the variance in both groups of students. These findings suggest that, despite meeting the language criteria at entry to university, the level of English with which Chinese L1 students begin their studies does not allow them to perform to their full abilities.

The above review showed that it is difficult for non-native speakers to compete academically with native speakers when their language is still developing. It was also shown that despite arriving at universities cognitively mature and with English L2 proficiency in place, international students' language skills are still significantly lower when compared to native speakers of English. Therefore, it is very likely that international HE students need to catch up linguistically with native speakers of English to be able to compete with them academically. However, evidence that could support this claim is missing.
The studies from the HE context suggest that the process of closing this gap, if possible, can be a very long process, but direct evidence is lacking. There are only a few studies that compared the language skills of home and international students at the start, and only one to date showing how the skills develop over one year. Therefore, international students' language needs to be studied not only to evaluate the magnitude of initial differences in a heterogeneous HE environment, but also to find out if the initial differences disappear in a timely manner allowing international students to achieve their full potential. In relation to undergraduate students, a time span of at least one year needs to be tracked, as this is the transition period where international students can prepare linguistically before the second year when they start earning grades that contribute to the final degree classification. The only piece of evidence suggesting a failure to close the gap after one year (Trenkic & Warmington, 2019) needs to be treated with caution, as the international students in this sample were L1 speakers of Chinese, a language very different from English, and it is difficult to generalise these findings to other cohorts of international students.

Despite some evidence, language is still called into question as the factor responsible for attainment differences. The argument is based on the grounds that international students meet the strict language requirements set by the universities (Crawford & Wang,

2015). However, findings from Trenkic and Warmington (2019) showed that language and literacy skills were predictive of academic attainment in Chinese L1 but not British students, which means that there exists a threshold of proficiency that must be met by international students. The average IELTS level of 6.92 demonstrated by the international students in this sample is not sufficient, which means that the cut-off score set by a university is not always aligned with the linguistic demands of the course. The claim undermining the role of proficiency in English in academic achievement is evaluated further in the next section, which discusses language-related entry requirements and the rationale behind setting the minimum entry requirements.

2.3.4 Linguistic requirements for English L2 students at English medium universities

Proficiency in English is crucial for EFL students as they need to understand lectures, communicate in seminars, comprehend tasks and assignments, read to acquire new knowledge, and participate in written examinations (Andrade, 2006). Even though international students reach the minimum language entry requirements set by universities, their language skills do not allow them to fulfil their full potential because they do not seem to be aligned with the linguistic demands of the courses (Trenkic & Warmington, 2019). This section provides more evidence supporting the claim that the level of English set by universities is not always sufficient to compete against native speakers of English. This is because international students still struggle linguistically after enrolment. The two main entrance proficiency tests, the International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOEFL), will be briefly described first, and some issues related to their implementation will be discussed.

IELTS is one of the standardised measures of proficiency recognised as an entry requirement for universities, mostly in the UK, Australia, and New Zealand. According to the test developers, it has been designed to assess 'whether you are ready to begin studying or training in an environment where English language is used' (IELTS Partners, 2017:10). The results of IELTS are expressed on a 9-point scale for each of the four components: reading, writing, speaking, and listening. The overall band score is the average score of all four language skills. The bands range from non-user (band 1.0) to expert user (band 9.0), where a

band score of 6.0 indicates a competent and 7.0 a good user. There is no pass mark on the test, which means that universities set their own cut-off scores. IELTS is regularly scrutinised by the research community in validation studies that provide guidelines for universities on how the cut-off scores should be set. These are usually band 7.5 for linguistically demanding academic courses and band 7.0 for linguistically less demanding courses (IELTS, 2019).

TOEFL is more popular in the USA and Canada, and similarly to IELTS, its score indicates whether the test taker obtained language proficiency that enables them to approach academic work in English in a meaningful manner. The original paper-based test, TOEFL (PBT) version, comprised the following areas: 1/ Listening section, 2/ Grammar and written expression section, and 3/ Reading comprehension section. The score on this test ranged on a scale between 0 and 677 points until the introduction of the Internet Based Test (TOEFL iBT) in 2006. Since that time, the rebranded IELTS test comprised of the four main parts testing reading and listening that are assessed through multiple-choice questions, and speaking and writing that required open-ended responses marked by trained professionals. Each of the four sections (reading, writing, speaking, and listening) has its own subscale with a maximum score of 30. The total TOEFL score is calculated by adding the points obtained in each of the four subtests (Sulistyo, 2009). The overall TOEFL iBT score is reflected on a scale between 0 and 120. Just like IELTS, this is not a pass–fail test and the cut-off scores are set individually by universities. The minimum required level of TOEFL varies significantly across different universities in the USA, with a range in cut-off scores from 70 to 100 obtained in the latest version of TOEFL iBT (Ginther & Yan, 2018).

As mentioned above, the cut-off scores in these tests are set by the HE institutions, as adequate proficiency depends on the context of study and other proficiency requirements. Following the IELTS developers' guidelines, higher scores are usually set for students within the field of humanities and social sciences (Edwards et al., 2007). Lower cut scores are set for science subjects such as mathematics or physics because they are linguistically less demanding and involve numerical operations and symbolic modes of working (Barton & Neville-Barton, 2003). However, the rationale behind setting these cut-off scores is not always clear, and they can differ from one institution to another and within a single institution as well. For example, it is a common practice that students who do not meet the minimum language requirements are offered a place on the condition that they complete a

preparatory course lasting a couple of weeks prior to the start of their courses. In one of the most recent studies (Lam et al., 2021), it was found that '[c]hanges to the minimum requirements are often motivated by recruitment considerations, benchmarked against rival institutions or neighbouring departments, but with limited engagement with guidance from IELTS' (p. 5). Despite the suggested cut-off scores of 7.0 and 7.5, lower-ranked universities in the UK accept students with overall IELTS scores as low as 5.5.

These admission practices further undermine the strength of the claim that international students arrive in English-medium universities with adequate skills in English. In many cases, it is very likely that the cut-off scores are not aligned well with the linguistic demands of the courses, and the language level with which international students begin their courses is an obstacle while pursuing their degrees. This last claim finds strong empirical support in studies demonstrating that, despite meeting the English criteria set by the HE institution, international students still struggle linguistically after enrolment at university. The scale of this problem has been recognised in universities in Australasia that have developed special provisions for those who, despite meeting language requirements, are still at risk of failing their course because of an insufficient level of English at the start of university. These include the Diagnostic English Language Needs Assessment (DELNA, Elder & von Randow, 2008; Read, 2008), Post-Enrolment Language Assessment (PLA, Murray, 2010), and Measuring Academic Skills of University Students (MASUS, Paton, 2007). The common goal of all these programmes was to target students who may be at risk of failing their degree programmes because of low proficiency in English. Each of them will be briefly discussed to demonstrate that many international students are at risk of failing their courses because of inadequate skills in English.

In New Zealand, students engaged in DELNA usually have difficulties with passing the screening part of the assessment (Elder & von Randow, 2008). The screening part assesses vocabulary knowledge and the speed of reading on a scale between 0 and 100. A score below 70 means that a student needs to take the second, diagnostic part of the assessment. The diagnostic component consists of tests in listening, reading, and writing, and the score in this part of the assessment is reported on a 6-point band scale where band 4 represents 'severe risk of academic failure' and the top band 8 & 9 states that 'no support is required'. The screening part of DELNA proved to be a good predictor of skills in English, as there was a

correlation of .82 between the screening part score and the overall band score on the diagnostic component. The results also showed that 93% of those who scored below the cut-off score on the screening component also scored below band 7 on the diagnostic assessment and were identified as at-risk students. Only a small percentage of those who performed below the cut-off on the screening were needlessly taking the more complex and time-consuming diagnostic assessment.

Post-enrolment screening has also been developed in HE in Australia, where it was concluded that 'current English language screening regimes are somehow inadequate in that students are gaining entry to higher education who subsequently struggle to meet the demands of their programmes due to weak language skills' (Murray, 2010: 346). The solution – similar to the one implemented in New Zealand – was the introduction of Post-Enrolment Language Assessment (PLA). Even though it would not prevent those students with weaker language skills from enrolling in the degree programme, it would help with identifying at-risk students early enough post-enrolment and providing them with relevant support. One of the greatest advantages of PLA is that this diagnostic test is not available publicly (unlike IELTS and TOEFL), which means that students cannot practise it. It also allows assessment of students holding a variety of qualifications, including both standardised tests and non-standard qualifications such as pathway programmes and pre-sessional courses. The authors list numerous benefits of PLA, such as introducing an even playing field for all assessed students and reducing the burden for the university caused by low levels of language skill (thanks to directing students to appropriate support straight away). These benefits are claimed to outweigh the costs related to discrimination on the basis of international status, the stigma of being identified as 'at risk', or logistical issues.

Another post-enrolment tool, Measuring Academic Skills of University Students, was developed at the University of Sydney. MASUS is a diagnostic language test to measure students' writing ability (Paton, 2007). It was developed in response to the university's concern that despite high levels of qualifications on entry, the level of attainment was not high, and language was identified as a possible cause of this. MASUS was intended to assess 'students' ability to write about a given body of knowledge in a reasoned and critical way, together with their ability to use the language resources appropriate for the required task'

(Paton, 2007: 103). The assessment was marked on a 4-point scale on each of the four criteria: information processing, academic English, grammatical correctness, and structure & development. The data accumulated on the MASUS results and students' academic achievement demonstrated that 25% of the undergraduate students in the Faculty of Economics and Business had unsatisfactory language skills and that 80% of that subset were international students. The data from postgraduate students showed that 88% of participants (*N* = 278) needed some language-related help. Only 17% of postgraduate students were native speakers of English, and 239 students spoke one of the Chinese dialects. The author concluded that 'English language competence is the fundamental issue for success at Australian universities, especially at the sentence level, because of the "fineness of meaning" necessary to academic discourse in whatever language' (Paton, 2007: 107).

Several factors may be responsible for the fact that international students struggle linguistically after enrolment at university. The most likely problem, highlighted above and discussed by Murray (2018), is that gatekeeping tests such as IELTS and TOEFL 'are being misused in that entry thresholds are being set too low, or that the tests are not really measuring what they need to be measuring or are only measuring part of what they need to measure' (p. 6). This final point may refer to the type of task; for example, the IELTS written tasks are criticised for a lack of authenticity in the academic context (Moore & Morton, 2005). Yet another problem connected to the standardised tests is that they may not reflect students' knowledge as a result of course preparation practices that train students in test-taking abilities rather than focusing on improving their general English skills (Hu & Trenkic, 2019). Another reason may be that some students arrive with results for non-standard language tests rather than IELTS and TOEFL, and admissions are unable to estimate the cut-off points for these tests correctly (Coley, 1999). It is also possible that many international students enter universities via alternative routes without the necessity to provide evidence of fluency in English, such as foundation and access courses, or that they graduate from English-medium high schools and this exempts them from providing qualifications in English (Fox, 2005). Some educators are also concerned that language proficiency requirements are sacrificed to attract students who are the main source of revenue for many Western universities (Jenkins & Wingate, 2015).

This section has shown that, despite meeting the language requirements set by universities and being admitted to university, many students still struggle linguistically. The likely reason is a mismatch between entry requirements and course demands. This may be unintentional as admissions are not always able to correctly assess the non-standard qualifications of international students, or it may be intentional in situations where universities set lower cut-off scores to secure a safe number of enrolments. Even if the level of English that international students arrive with allows them to pass their courses, it is not always sufficient when compared to the abilities of native speakers of English introducing an unequal level playing field. An inadequate level of English from the start may mean that international students cannot minimise the magnitude of linguistic difference in a timely way and this in turn may cause under-attainment in relation to native speakers of English. Despite the large body of studies on differences in academic achievement between English native speakers and international students, several researchers did not find any difference, which gives rise to some further questions on the role of English in academic attainment. This issue will be discussed in the next section.

2.3.5 Attainment differences are not always found. What does it mean?

Despite strong evidence on attainment differences between international and home students in English-medium universities, several studies have showed that international students are not always disadvantaged and that they can perform as well as the domestic students (Bers, 1994; Crawford & Wang, 2014; Hartnett et al., 2004; Isonio, 1994; Rankin et al., 2003). This may suggest that language is not always an obstacle while studying in English L2. For example, Rankin et al. (2003) investigated differences between home and international students' grades obtained at the end of the first semester in an introductory accounting course in an Australian university. The results showed no significant difference in grades between native speakers of English and English learners. In another study, also investigating accounting students but in a UK context, Crawford & Wang (2014) obtained similar results. This study investigated four cohorts of undergraduate students in four consecutive years from 2006–2007 to 2009–2010 and no significant differences were found after the first semester between home and international students. Similarly, Berman &

Cheng (2001) found that undergraduate international students obtained better grades than their English L1 counterparts after the first semester at a Canadian university.

International students (*N* = 271) also obtained significantly greater GPAs when compared to home students (*N* = 442) enrolled in a three year diploma in Singapore (Nasirudeen & Xiao, 2020). The international student sample came from Asian countries where English is not the language of instruction at school, mostly from China, Myanmar, Indonesia, Vietnam, and Malaysia. The results showed that despite only a small difference in the mean scores, international students obtained significantly greater GPAs when compared to the home students. The conclusion based on the results was that international students can perform as well as, or even better than, home students. However, it is worth pointing out that the home students in Singapore are very different linguistically from home students in English-speaking countries. They are part of a multilingual society speaking a variety of L1s, with English being the lingua franca and the language of instruction at school. The fact that home students were schooled in English-medium instruction could still have some positive influence on the home students' language and academic performance. According to the questionnaire administered to both groups, the group of international students perceived the four skills in English as more difficult when compared to home students. Furthermore, the correlations performed between self-perceived difficulties with language and GPAs showed that international students struggled with more language areas when compared to their home counterparts. International students' responses to 25 (out of 40) survey items correlated with their GPAs, whereas home students' responses to only 7 survey items correlated with their GPAs.

The above review shows that international students can perform as well as home students; however, the common denominator in all the studies that found no difference between the two cohorts in question is that the attainment was measured early on while at university, usually after the first year, or even the first semester (Berman & Cheng, 2001; Crawford & Wang, 2014; Rankin et al., 2003). These findings comprise only one part of a bigger picture and in fact fit some global trends discovered in longitudinal studies where academic achievement was tracked for several years. For example, Crawford & Wang (2015) found that Chinese students outperformed home students at the end of the first year, but then

they lost the competitive edge and their performance in the consecutive years worsened in relation to home students. They also graduated with lower degree classifications.

The same pattern was observed in another longitudinal investigation of attainment differences between home and international students in Canada. Using business students' graduation averages, He & Banham (2009) found that international students outperformed home students, but only in the first year. After the second year at university, the international students experienced a dramatic drop in their mean scores, and their performance in the consecutive years was also lower when compared to home students. Similarly, Patkowski et al. (1997) compared cumulated GPA semester by semester over a period of three years at two types of HE institution in the USA. The EFL students were divided into three levels of linguistic abilities: low, medium, and high. According to the results, students at low and medium levels of English performed competitively with native speakers of English at the end of the first semester at university, but a reverse trend was observed in the final year, with native speakers of English obtaining higher GPAs.

It is also important to point out the methodological shortcomings in studies that did not find differences in grades when comparing home and international students, specifically in relation to the linguistic characteristics of the students taking part in research studies. For example, almost 80% of the EFL sample in Crawford & Wang (2014) obtained A-level qualifications or equivalent in the UK, with the following 8% studying for an International Baccalaureate with English as the language of instruction prior to university, making them an atypical international cohort. Also, the group of home students in Nasirudeen & Xiao (2020) was not typical in the sense that home students were in fact multilingual students speaking English as a lingua franca. This methodological issue can obscure the findings obtained in these studies. Another shortcoming is that academic attainment was compared very early on while at university, which does not allow generalisation beyond one year at university. This is because neither the demands of the course nor language skills stay at the same level as at the time of entry to university; rather, they are both subject to change over the course of study. Given that the content at university becomes more difficult from one year to another, and that there are no longitudinal studies that show the development of English skills in international students, it is difficult to predict whether students would maintain the same level of attainment throughout the whole duration of their studies.

Therefore, findings based on achievement differences obtained after the first semester at university cannot guarantee that this pattern would persist.

This section has shown that, even though some studies find a lack of difference in attainment, this does not necessarily mean that language is not important in academic achievement. These findings are rather due to some methodological limitations. One of the limitations of the existing studies is that they look at attainment very early on, usually after the first semester at university, which further justifies the need to study differences in attainment and language and its development longitudinally. Another limitation of studies that do not find a difference is the use of atypical samples of home or international students without clearly defining, describing, and controlling them for a variety of language-related variables. Therefore, there is a need for more rigorous studies in terms of participant selection and controlling for variables that may obscure the patterns in findings.

2.3.6 Summary so far

This section has shown that the number of international students arriving at British universities has led to the development of a linguistically heterogeneous environment. This brings benefits for everyone: the receiving institutions, the wider economy, and the whole student body. To preserve these benefits, it is necessary to sustain a steady international enrolment. This can be achieved if international students are provided with conditions that allow them to compete linguistically against home students who study in their first language. It is therefore crucial to compare the linguistic abilities of these two cohorts at the start of their degrees because all students are assessed using the same criteria when working towards their degrees. Despite the great value attached to degree classifications, international students obtain fewer good degrees, which may jeopardise their educational and professional goals, and the reasons are poorly understood.

Research shows that second language skills constitute one of the most important factors in academic achievement. At the same time, the level of English with which international students begin university is significantly lower when compared to the native speaker. Given the theoretical underpinning from limited-proficiency children, it is very

likely that international students cannot compete with the ENS until they catch up with them linguistically, but evidence in the context of HE is lacking. It is therefore of vital importance to compare language skills in international and home students and to track their development over time as there is no evidence on how long it takes for international students to catch up linguistically and to become equal competitors in environments where people with varying degrees of English knowledge study under the same conditions and towards similar objectives and goals. The findings would in turn shed light on the reasons for the attainment differences between these two cohorts.

Language is still questioned as the potential factor that can impact academic attainment. Despite the claims that international students' language is adequate because they meet the required language criteria, international students struggle linguistically after enrolment. The most likely reason is that the language requirements set by universities are, for various reasons, not well aligned with the linguistic demands of the courses. Also, studies that compared academic achievement and found no difference in attainment are methodologically flawed and make conclusions that cannot be generalised beyond year one of university study. To understand the reasons behind attainment differences, more research needs to be done that investigates the actual language and uses a more comprehensive and rigorous methodology. It would ideally compare not only language and literacy skills but also phonological skills, as they have proved to be an important factor in academic achievement. If language impacts academic attainment, we need to know which specific skills in English are the most important at university, how quickly they develop, and whether international students can catch up over the period of one year. The next section describes the importance of language in academic achievement and the importance of vocabulary, reading, and writing, which gives rise to their assessment in this study.

2.4 English L2 as a predictor of academic achievement

The increase of internationalisation and linguistic diversity has stimulated research in cross-cultural differences, with language being the top factor studied within this strand. Second language skills are crucial at university as students need to communicate in seminars and with their tutors, complete assignments, and take examinations. All these activities have a direct impact on academic achievement, making language one of the most important

predictors of study success. In addition to language, there exist many other factors affecting attainment, which can be grouped into academic, psychosocial, cognitive, and demographic categories (see Li et al., 2010, for an overview). Prior academic achievement, learning skills, habits, strategies, and approaches are factors that can be listed in the academic category. Some of the factors that belong to the psychosocial dimension are social integration, financial situation, motivation, social and emotional support, and psychological health. The cognitive factors include self-efficacy and attributional style, which can be defined as the student's attitude, e.g., pessimistic or helpless. Finally, the demographic factors include aspects such as age and gender. Therefore, when researching academic achievement from any given perspective, it is important to bear in mind that academic achievement is a construct that can be impacted by many, sometimes interrelated factors.

The present study investigates academic achievement from the linguistic angle. This section will review some key studies on the importance of English in academic achievement, focusing mainly on two strands of studies. The first one comprises literature on the predictive role of IELTS and TOEFL in academic achievement and the second strand looks specifically at different types of knowledge and literacy skills. Because language is not independent of cognitive abilities, there will be a short section showing that intelligence and working memory, cognitive abilities that underpin language, can be direct predictors of academic attainment.

2.4.1 IELTS and TOEFL and academic achievement

Research on the predictive validity of IELTS (and TOEFL, which will be discussed later) usually explores correlational analyses, with test scores being correlated with students' GPAs. The correlations, however, are in most cases weak and the findings are not always consistent, which has called into question the role of English as a predictor of study success. This is because some key studies show a wide range of correlation coefficient values: r = .06 (Fiocco, 1992), r = .10 (Oliver et al., 2012), r = .31 (Ferguson & White, 1998), r = .35 (Elder, 1993), r = .40 (Woodrow, 2006), r = .46 (Yen & Kuzma, 2009), r = .49 (Al-Malki, 2014), r = .52 (Bellingham, 1995), and r = .54 (Hill et al., 1999). Some of the studies found correlations that were non-significant (Kerstjens & Nery 2000; Ingram & Bayliss, 2007) or even negative (Cotton & Conrow, 1998). However, these studies reveal a consistent pattern, showing a

relationship between test scores and academic achievement at lower levels of proficiency (Bellingham, 1995; Elder, 1993; Feast, 2002; Ferguson & White, 1998; Hill et al., 1999; Oliver et al., 2012; Woodrow, 2006).

For example, Woodrow (2006) found that language is a predictor of academic success before the threshold of proficiency is reached and in the context of her study this threshold may correspond to an IELTS score of 6.5. The data came from postgraduate international students enrolled in social science subjects (*N* = 62). In another study, Bellingham (1995) found that an IELTS score of 6.0 constituted a threshold that could predict study success, as students below it had only a 20% chance of obtaining a pass and the success rate of those obtaining a score of 6.0 or higher increased to 50%. This study was based on participants (*N* = 38) enrolled in a two-year National Certificate of Business programme in New Zealand.

In a similar vein, Ferguson & White (1998) concluded that English skills at entry determined the probability of participants failing their course and that IELTS is more predictive for those at lower levels; the threshold was set at band 6.0. The participants in this study comprised a heterogeneous sample of students from Asia, Africa, the Middle East, Europe, and Latin America (*N* = 28) in master's courses in science at the University of Edinburgh.

The IELTS score was also important at a lower level of proficiency in a study by Elder (1993), but the threshold of proficiency was set at a much lower level – an IELTS score of 4.5. The data came from 32 candidates admitted to the teacher's diploma course in six educational institutions in Melbourne. In another study from Australia, students (*N* = 35) were grouped into four levels according to their language ability as expressed by their IELTS score: group 1 with IELTS 6.0, group 2 IELTS 6.5, group 3 IELTS 7.0, and group 4 IELTS 7.5–8.5 (Hill et al., 1999). The results showed that the GPA increased with an increase in language ability, with the largest increase in the highest language ability group. The difference in scores between the highest ability group and the other three groups proved to be statistically significant. In sum, despite mixed results obtained in IELTS correlational studies, language seems to be important, especially for those at a lower level of proficiency, and the threshold of proficiency usually seems to correspond to an IELTS band between 6.0 and 6.5.

Although TOEFL is more popular in the USA and only a handful of students taking part in the present study demonstrated this qualification, it is worthwhile to look at research in this context because it seems to mirror the findings obtained with IELTS. Studies on the predictive validity of TOEFL for academic achievement, similarly to those on IELTS, show inconsistent results. Different studies have found correlations ranging from weak (Cho & Bridgeman, 2012; Light et al., 1987) to moderate and strong (Al-Musawi & Al-Ansari, 1999; Ayers & Peters, 1977; Johnson, 1988; Sharon, 1972; Vinke & Jochems, 1993). Some studies found no relationship (Ayers & Quattlebaum, 1992; Krausz et al., 2005; Hwang & Dizney, 1970) or even negative correlations (Bridgeman et al., 2015; Neal, as cited in Graham, 1987; Neal, 1998). However, these studies confirm that language plays a crucial role at lower levels of proficiency in English and that the probability of obtaining high marks increases with an increase in TOEFL score at entry to university. Higher TOEFL scores are associated with a higher level of achievement until a threshold in language proficiency is reached (Cho & Bridgeman, 2012; Graham, 1987; Johnson, 1988; Light et al., 1987; Vinke & Johems, 1993; Yul & Hoffman, 1990).

In one of the first studies on the predictive validity of TOEFL, the author suggested that language is the key, especially for students at lower levels of proficiency (Graham, 1987). This conclusion was made after surveying studies with participants at different proficiency levels. For example, in a study by Gue & Holdaway (1973), the students obtained TOEFL scores of 424–447, and in another study by Light et al. (1987), the sample was at a much higher proficiency level (400–677). TOEFL was a good predictor in the study involving participants at the lower level of proficiency, where the correlational coefficient reached r = .49–.59. A closer look at the study by Light et al. (1987) seems to confirm these conclusions. The participants (N = 387) were ordered according to their TOEFL level into 5 groups: 400–529, 530–549, 550–569, 570–599, and 600–680. The results showed that those at a higher level of proficiency obtained a greater number of credit hours, which is the unit of measuring educational credit usually based on the number of classroom hours per week in a term. This effect was strongest for those with TOEFL levels of 530–549 and 570–599, and it started to level off at a TOEFL score of 599.

The study by Light at al. (1987) was replicated by Johnson (1988) and the author concluded that ‘the lower the language proficiency, the more important the role it plays in academic

success' (Johnson, 1988: 165). This study investigated a group of undergraduate international students (*N* = 196) in a US university. The sample was divided into two groups according to whether their TOEFL score was below or above 500. The analysis showed that those with TOEFL scores below the cut-off of 500 obtained significantly lower GPAs than those with TOEFL scores at the higher end. In another study by Yul & Hoffman (1990), academic achievement was conceptualised as a positive recommendation for the position of teaching assistant. The students (*N* = 233) who were granted a positive recommendation had on average significantly higher TOEFL scores than those who were not granted a positive recommendation, $t(231) = 9.34$, $p < .001$. A further multiple logistic regression analysis showed that the TOEFL score had a significant effect on the recommendation status in such a way that each one-point increase in the TOEFL score corresponded to a 3.6% increase in the odds of being recommended to teach.

Vinke & Johems (1993) also found that language skills were the best predictors of academic achievement at a lower level of proficiency. Correlations between TOEFL score and academic achievement in students grouped according to a TOEFL cut-off score of 540 showed that students with lower TOEFL scores (*N* = 49) obtained significantly lower marks than those with higher TOEFL scores (*N* = 41), $t = 4.63$, $p < .002$. In addition, the overall success rate of all students was 84.4%, and this pass rate increased to 97.6% in the more proficient students and dropped to 73.5% in the less proficient subgroups. These studies show that, at least in the older version of TOEFL, the threshold of proficiency for students was placed somewhere between 450 and 550, but it is difficult to translate these scores to the new scale of TOEFL iBT.

The first conclusion based on the above findings is that studies on the predictive validity of IELTS and TOEFL bring mixed results in terms of the strength of the correlation coefficient. The interpretation of correlation coefficients in this context is especially difficult because the predicted variable (academic achievement) depends on many factors (e.g., academic, psychosocial, cognitive, demographic, and linguistic factors), but it is very unlikely that a correlation coefficient value close to 1 will be found. In this case, the strength of correlation should be interpreted in light of the multitude of factors that may affect academic achievement. If a construct is expected to be predicted by many factors, then a correlation coefficient of .30 can be interpreted as very strong. Therefore, weak correlations observed

in the above-mentioned studies do not mean that the relationship between second language skills and academic achievement does not exist, but that there are other factors involved, and a relatively weak correlation can be very meaningful.

Another explanation behind the weak correlational coefficients is the fact that there is very little variability in the proficiency test scores (Pearson, 2021). If IELTS or TOEFL scores were all the same at entry, 'the correlation between these scores and any measure of study success will automatically be zero' (Daller & Phelan, 2013:177). Cho & Bridgeman (2012) also emphasise that 'when the data do not represent a full range – that is, there is no information regarding how those who did not get selected would have performed afterward – a correlation is underestimated' (p. 425). These claims find support in a study by Ginther & Yan (2018), who corrected their correlations for the restricted range of scores. Their results showed that the correlation between the overall TOEFL score and the GPA for the first year at university in the first sample (*N* = 740) was .07. However, the adjusted *r* value increased to .21. A similar trend was observed in another cohort of students (*N* = 554). The correlation between the overall TOEFL and the GPA for the first year at university increased from .04 to .16 after correcting for the restricted range of scores. These findings seem to be consistent with correlational studies where students demonstrated a wide range of scores upon entering university. For example, Bellingham (1995) explains that the relatively high correlation coefficient in their study (*r* = 0.52) was obtained because of the wider range of IELTS scores at entry. The same was concluded by Kerstjens and Nery (2000), in whose study the range of IELTS marks was very wide, between 3.5 and 6.0. These findings show that low correlational coefficients do not necessarily mean that there is no relationship between language and attainment, because the correlations can be distorted by truncated sample sizes.

There is further evidence suggesting that even weak correlational coefficients may be very meaningful. In a study by Cho & Bridgeman (2012), the authors conducted two types of analyses: correlations and contingency tables. The IELTS iBT and GPAs came from undergraduate and graduate students (*N* = 2,594) at 10 different institutions. The correlations for all undergraduate students (*r* = .19), as well as for graduates (*r* = .16), proved to be weak. However, the alternative analysis using contingency tables showed that students with higher entry level scores tended to have higher GPAs. The authors concluded

that ‘even small correlations or seemingly trivial amounts of variance explained may be an indication of a meaningful relationship between two variables’ (p. 439). This means that the correlations are low not because language does not matter, but because there are other reasons explaining the findings – such as the multiple factors that can impact academic achievement, or truncated scores in proficiency tests such as IELTS and TOEFL as a result of students beginning their studies as soon as they obtain the minimum cut-off score that allows them to enter university.

In addition to weak correlational coefficients and difficulties with their interpretation, another conclusion based on studies on the predictive validity of IELTS and TOEFL is that language is a better predictor of academic success at lower proficiency levels. However, this does not necessarily mean that language is not important at higher levels of proficiency. Second language skills are a necessary but not sufficient condition for academic success. There is a threshold of proficiency in English that students need to meet, and once this is achieved, the importance of English language proficiency can be suppressed by other factors (Graham, 1987). What is more, there are some methodological shortcomings that may prevent detecting the importance of language at higher levels of proficiency (see Pearson, 2021, for an overview). Research studies on the predictive validity of IELTS and TOEFL define academic achievement in different ways (Cotton & Conrow, 1998; Ferguson & White, 1998), measure it at different points in time (Elder, 1993; Light et al., 1987; Yen & Kuzma, 2009), and vary as to whether the correlations involve the overall test score or subscale results (Bridgeman et al., 2015; Ginther & Yan, 2018). Researchers do not always control for extraneous variables that may obscure the pattern of results, such as the subject studied or participants’ L1s. These two variables are discussed briefly below, as they are relevant for the design of the present study.

One of the variables that can impact the results is the subject studied at university. This is because not all university courses are the same in terms of linguistic demands. In general, subjects within the humanities and social sciences are linguistically more demanding than subjects in science (Woodrow, 2006). It is generally more likely that a positive relationship between proficiency and academic outcome will be found when the area of study is controlled for (Bellingham, 1995; Elder, 1993; Ferguson & White, 1998). For example, Light et al. (1987) found differences in correlation strength among different subjects, and that in

subjects involving quantitative skills such as science, English is predictive to a lesser extent. Bridgeman et al. (2015) also found that correlations differed depending on the subject area, whether it was business, engineering, arts, or science. Johnson (1988) found that the subject studied was an important factor, as students with the lowest TOEFL scores and studying linguistically less demanding courses performed better than those who were at the same proficiency level and studying subjects that were more demanding linguistically. Cho & Bridgeman (2012) showed that when a sample was divided according to the subject studied, higher correlation coefficients were obtained for postgraduate students than when the whole sample was analysed. TOEFL also proved to be a better predictor in subjects that are linguistically more demanding (humanities and arts, social science) than in science and engineering. Therefore, studies investigating language and academic achievement should strive for samples at least within the same faculties.

In the predictive validity studies, participant samples consist of students with different L1 backgrounds who may have very different L2 learning opportunities. However, participants' L1 is not always controlled for in the IELTS and TOEFL studies. The importance of students' L1 was especially pronounced in a study by Bridgeman et al. (2015). When the sample studied was controlled for nationality, the relationship between IELTS scores and GPAs proved to be very strong. The data came from US undergraduate university students enrolled in a range of subjects. The participants (N = 787) comprised international students from China (40%), India (16%), Vietnam (8%), and Korea (5%). The correlation between TOEFL scores and combined grades from three semesters in the whole sample of students was significant but weak (r = 0.18). However, when the sample was more uniform with participants grouped according to the subject studied and country of origin, a different picture emerged. The correlation between TOEFL scores and academic grades in Chinese L1 students in engineering programmes (N = 64) increased from r = .18 to r = .58, and further to .77 when it was corrected for the restricted range. These results show that a well-selected sample (controlled for the subject studied and for nationality) can yield a meaningful correlational coefficient between proficiency test and GPA. Single-nationality correlational studies seem to yield stronger and more consistent results in general (Ayers & Peters, 1977; Vinke & Johems, 1993; Yen & Kuzma, 2009).

The findings on the predictive validity of standardised tests are not always consistent, but the results do not necessarily mean that language is not important in academic achievement. In a situation where the outcome variable can be predicted by many factors, a relatively weak correlation may in fact be very meaningful. Research in this area also shows that language is less predictive at higher levels of proficiency but again, once the threshold of proficiency is attained, more factors may come into play. What is more, researchers need to control for variables that can potentially distort the pattern of results, with the subject studied and the participant's L1 being the most important ones. The next section discusses the relationship between vocabulary, reading, writing, and academic achievement and the findings are much more consistent there, which further demonstrates how important second language skills are at university. The discussion will include consideration of some research-related challenges as they are relevant to the methodology used in the present study.

2.4.2 Vocabulary and academic achievement

There are many aspects of vocabulary knowledge that need to be considered in research and these include defining what a word is, sampling words for testing, deciding on the degree of knowledge of a word (vocabulary size vs vocabulary depth), and selecting an appropriate tool for testing. There are several ways of defining what a word is. This can be as a *token*, each word form in a spoken or written language; a *word type,* each different word used in a text; or a *lemma*, which consists of a headword and its inflected forms belonging to the same part of speech, so that *paint* (N) and *painter* (N) would constitute the same lemmas, but *paints* (V) would be considered as a different one. Finally, there is the concept of a *word family*, which consists of a base form and related inflected and regularly derived forms (Nation, 2001). For example, *build*, *built, builder*, *building* all belong to one word family. In researching vocabulary knowledge, the way a word is defined should be aligned with the learner's proficiency level and the associated learning burden. For example, knowledge of a word family requires a broad knowledge of inflectional and derivational forms, perhaps involving multiple affixes, and also the ability to apply this knowledge (Brown et al., 2020). It may be more suitable in the context of more advanced speakers who have some control of word-building devices and who can identify relationships between

regularly affixed members of a family. It is also claimed that inflectional and derivational morphology develops alongside general proficiency (Mochizuki & Aizawa, 2000), further supporting the suitability of defining a word in terms of a word family at higher proficiency levels.

Almost all available vocabulary tests are based on the Frequency Model (Milton, 2009), according to which words that are more frequent in a language are more likely to be known than words that are less frequent. The words used for vocabulary assessment are therefore sampled from a variety of word lists and different frequency bands. These lists are based on dictionaries or large corpora and rank words in order of descending frequency. They are usually organised in bands of 1,000 words at a certain frequency level, with around 10 words tested from each frequency level. It is important to note that frequency is the primary factor but only one of many that can affect vocabulary acquisition. For example, Hashimoto (2021) claims that the frequency of a word does not always correlate with the likelihood of a learner knowing the word and that there exist other factors that predict word knowledge, including word length, polysemy, concreteness, and cognateness. However, frequency is the predominant factor utilised in almost all available vocabulary tests, and more instruments exploring different factors need to be developed and validated for research purposes to determine their importance.

Another aspect of vocabulary knowledge measurement involves vocabulary size and depth. The former tells how many words one knows whereas the latter tells how well a word is known. Research shows that these two aspects of vocabulary knowledge are highly related (Shimamoto, 2000; Schmitt, 2014; Vermeer, 2001). For example, Mehrpour et al. (2011) investigated the relationship between these two dimensions of vocabulary knowledge (size and depth). The data obtained from 60 EFL learners using two different vocabulary tests showed a high level of correlation between the two dimensions of vocabulary. Similar conclusions were reached by Qian (2002), who found that even though vocabulary depth was a better correlate to some aspects of knowledge of English, the two dimensions of vocabulary knowledge were also related to each other to a great extent, and both are good predictors of reading skill. Finally, vocabulary can be tested in receptive or productive vocabulary tests. The productive use of vocabulary is expressing a meaning in speaking or

writing by retrieving the word from memory and producing it in speaking or writing. The receptive use of vocabulary, on the other hand, consists of recognition of a word when one hears or sees it. Receptive vocabulary is greater in size than productive vocabulary because it may include words that are only partially known. This is important, especially in reading, as partial or superficial knowledge of a word aids in guessing its meaning from the context (Laufer, 1998).

L2 vocabulary learning targets are usually discussed in the context of vocabulary knowledge in the native speaker. Research shows that an educated native speaker of English knows on average 17,000 word families. The rate of vocabulary acquisition in this population is estimated to be around 1,000 word families per year but there are a lot of individual differences and learning gains may differ from one person to another (Goulden et al., 1990; Zechmeister et al., 1995). It was shown above that there are significant differences between ENS and EFL in their vocabulary size at the start of their university degree (see section 2.3.3). However, when it comes to the rate of vocabulary acquisition, English learners can reach comparable learning rates to ENS while in the second language environment. This was found in a study by Milton & Meara (1995) where 53 European L1 students at an advanced proficiency level were tested using the Eurocentres Vocabulary Size Test. The results showed that while learning English in the second language context, the average growth in each person was comparable to the growth estimated for a native speaker of English in adolescence.

Well-developed vocabulary is especially important in reading as the size of vocabulary can predict the level of text comprehension. Research shows that the 2,000 most frequent headwords cover between 80% and 85% of the words in many spoken and written texts (Nation & Waring, 1997; Nation & Newton, 1997), which means that on average 2 words in every line are unknown to readers at this proficiency level. According to Na & Nation (1985), this proportion of words is not sufficient to guess the meaning of the unknown words from the context and they suggest that coverage of at least 95% of the text is needed to be able to do that. Some recent studies suggest that the text coverage must be between 8,000 and 9,000 of the most frequent word families for unassisted comprehension of written texts and between 6,000 and 7,000 for spoken texts (Nation, 2006), which corresponds to around 95–

98% of text coverage (Hsuech-Chao & Nation, 2000). Knowledge of the 14,000 most frequent word families covers around 99% of a text and to read with minimal disturbance from unknown vocabulary, language users need between 15,000 and 20,000 word families. Therefore, the more vocabulary is known, the more text can be covered and this in turn leads to acquisition of more words.

Because of its importance in reading, vocabulary is a good predictor of reading comprehension (Beglar & Hunt, 1999; Laufer, 1992; Qian, 1999; Stæhr, 2008). This is because many unknown words could cause difficulties in text processing and, consequently, breakdowns in comprehension. A well-developed vocabulary is therefore crucial for reading in general and for reading academic texts specifically in the university context. For example, in a study with 92 first-year university students speaking L1 Hebrew and L1 Arabic, Laufer (1992) found correlations of $r = .50$ and $r = .75$ between two different vocabulary tests (Vocabulary Levels Test and Eurocentres Vocabulary) and reading comprehension. In another study including 30 undergraduate students at Bangkok University, Pringprom & Obchuae (2011) also found a relationship between vocabulary size and reading comprehension. The instruments used for assessment of vocabulary and reading comprehension were the Vocabulary Levels Test to assess receptive vocabulary size and a multiple-choice question assessment of reading comprehension. Rashidi & Khosravi (2010) investigated 38 senior university students and administered two vocabulary tests assessing vocabulary depth and breadth. The correlation between the Word Associates Test and reading comprehension was $r = .87$, and the correlation between the Vocabulary Levels Test and reading comprehension was $r = .75$. The findings suggest an interrelationship between the two measures of vocabulary. These findings show the importance of well-developed vocabulary when pursuing HE with English L2.

Knowledge of vocabulary can predict not only level of text comprehension, but also ultimate academic attainment at university (Alsager & Milton, 2016; Daller, Müller et al., 2021; Daller & Xue, 2009; Daller & Phelan, 2013; Elder & von Randow, 2008; Harrington & Roche, 2014; Masrai & Milton, 2017; Morris & Cobb, 2004; Müller & Daller, 2019, Murray, 2010; Read, 2008; Roche & Harrington, 2013; Szabo et al., 2021; Trenkic & Warmington, 2019; Yixin & Daller, 2014). Vocabulary proved to be a strong predictor of grades in international students

at a UK university (Daller & Phelan, 2013). The participants consisted mainly of students speaking European L1s (*N* = 74). Four tasks were administered: the C-test (a gap-fill with the second half of every second word deleted, see Daller, Müller et al., 2021 for an overview), Sigma V5 (a measure of verbal intelligence), and IELTS-like listening and writing tasks.

The written work was assessed for content and for lexical richness using the Guiraud Advanced measure that is based on the number of different words used in a text. Three tasks were administered at the start of university: C-test, verbal intelligence, and writing. Two of the tasks, C-test and listening task, were administered after one year to a subset of participants (*N* = 20), with the C-test being the only task administered at both time points. The results showed that individual measures can predict study success to some extent but a combination of three measures (the first C-test, Sigma, and Guiraud Advanced) explained GPA almost entirely in one of the analyses (R^2 = .958). Another analysis showed that a third of the variance in students' grades can be predicted based on these three measures of vocabulary (R^2 = .372). In sum, vocabulary profiles alone explained between 33% and 96% of the variance in students' marks in this study.

The findings obtained by Daller & Phelan (2013) seem to be consistent with findings from another university in the UK but with students speaking a different first language. The knowledge of vocabulary proved to be predictive of academic achievement in a cohort of undergraduate and postgraduate students speaking Chinese L1 (Yixin & Daller, 2014). Two tasks were administered to students (*N* = 60) at two time points, at the start of university and one year later. These were the C-test explored by Daller & Phelan (2013) and a writing task adapted from IELTS practice test materials. Academic achievement was conceptualised as the GPA obtained at the end of the second semester at university. The written task gave rise to three measures of vocabulary: Guiraud (the type-to-token ratio), Guiraud advanced, and D (a lexical richness parameter that models the distribution of types and tokens in a text). The analyses included correlations between each of the 5 measures and GPA. At both time points, Guiraud yielded the highest and significant correlations with GPA (*r* = .526 at T1 and *r* = .493 at T2). This was followed by the C-test scores (*r* = .317 at T1 and *r* = .404 at T2). All five measures correlated significantly with GPA at T1 but at T2 no correlation was found between Guiraud Advanced, D, and GPA.

Vocabulary knowledge has been found to be an important predictor of grades also in the foreign language context. Two studies conducted in the Sultanate of Oman investigated the importance of vocabulary knowledge in academic achievement (Harrington & Roche, 2014; Roche & Harrington, 2013). A Timed Yes/No (TYN) vocabulary test was used in both studies and involves the test taker selecting known words from a word list. To control for guessing, the test taker is penalised for selecting non-words inserted amongst the real English words in the list. The test also yields a measure of the speed of answering which is a complementary measure of the knowledge of vocabulary. In one of the studies (Roche & Harrington, 2013), the TYN vocabulary test and an IELTS-like writing task were administered. Academic achievement was conceptualised as GPA. The participants consisted of first- and fourth-year university students who were L1 Arabic speakers (N = 70). According to the results, vocabulary size and speed correlated significantly with both the writing measure and GPA. When the two variables (vocabulary and writing) were considered together, they explained almost 25% of the variance in GPA (R^2 = .234).

Knowledge of vocabulary proved to be a significant predictor of academic attainment when yet another vocabulary test was used in a new learning context (Masrai & Milton, 2017). This study was conducted in Saudi Arabia and used the XK Lex vocabulary test, which is a measure of overall vocabulary size test based on 10,000 words with 10 words from each 1,000 frequency band. The test format resembles that in the timed Yes/No test as participants are presented a list of words and non-words and have to decide which words they know. In addition to this general vocabulary test, another vocabulary test administered in this study assessed academic vocabulary knowledge specifically. Academic achievement was conceptualised as GPA based on first-year grades. The participants consisted of undergraduate students in language and translation programmes (N = 96) in Saudi Arabia from two universities. Both vocabulary tests correlated strongly with students' GPA; the academic vocabulary test (r = 0.73, p < .001) and the general vocabulary test (r = 0.78, p < .001). These results show that both vocabulary tests predict a significantly large portion of GPA in a previously unexplored participant sample. The academic vocabulary predicted 53% of the variance in GPA and the general vocabulary a little less, 47% of the variance. The authors also concluded that both tests are assessments of the same construct, which means that either test can be used to test vocabulary knowledge in university students.

2.4.3 Grammar and academic achievement

In the traditional sense, grammar is 'primarily concerned with the well-formedness (or ill-formedness) of a sentence or a subpart of a sentence such as a clause or a phrase' (Shiotsu & Weir, 2007:106). Other scholars treat grammar more broadly; for example, Horrocks (1987) defines it as 'concerned with the principles according to which words can be combined to form larger meaningful units, and by which larger units can be combined to form sentences' (p. 24), and for Crystal (1997) grammar is 'the way in which words are arranged to show relationships of meaning within (and sometimes between) sentences' (p. 94). Purpura (2004) goes even further in defining grammatical knowledge as knowledge of phonological, lexical, and cohesive forms and their meaning.

The broad range of definitions of grammatical knowledge and the different ways of understanding it lead to some challenges in measuring it. For example, Shiotsu & Weir (2007) used the Test of English for Educational Purposes in their investigation of the importance of grammar in reading comprehension. This test consisted of sentences with one part replaced with a blank. Participants' task was to select the correct answer from among four possible options which had similar semantic content, but only one of which satisfied the syntactic constraints imposed by the sentence in question. The researchers' aim was to arrive at test items that were decontextualized and focused on the notion of acceptable sentence construction and less on sentence semantics. Other researchers have used multiple-choice and cloze items (Brisbois, 1995), gap-filling of a continuous text (Alderson, 1993), or a picture-matching format (Dąbrowska, 2018). However, one of the most widely used test formats is the grammaticality judgement test where participants are presented with a list of sentences and must select those they consider grammatical (Flege et al., 1999; Johnson & Newport, 1989).

Grammatical knowledge is an important factor in reading comprehension (Barnett, 1986; Haynes & Carr, 1990; van Gelderen et al., 2004), but the literature on this topic is very limited. One of the challenges in this strand of research is the way grammar is conceptualized and assessed. Researchers stress that instruments testing grammar and syntax should reflect the construct being measured. They point out that there is an overlap in the different constructs being measured and that what is meant to be a test of grammar

also reflects other abilities. For example, it is difficult to isolate vocabulary and grammatical knowledge. Also, grammatical tests used in research would involve the processing of visually presented text, which may be judged as assessing reading (Shiotsu & Weir, 2007). Alderson (2000) also agrees that it is difficult to isolate the importance of knowledge of particular syntactic structures, or the ability to process them, from some aspects of second language reading, but states that the 'ability to parse sentences into their correct syntactic structure appears to be an important element in understanding text' (p. 37). There is also a lack of research on differences in grammatical knowledge between native speakers of English and English learners at the start of university and this gap needs to be addressed.

Another challenge related to testing grammar and grammar development is the difficulty of assessing this construct in a native speaker of a given language. In general, it is very difficult to test for differences in grammatical knowledge in adult native speakers of a given language, because their grammar is well developed and they tend to perform at ceiling. What is more, testing grammatical knowledge among an adult, well-educated population is even more challenging because of their amount of exposure to printed text (Dąbrowska, 2018). The greater the exposure, the better the grammar tends to be. Therefore, it is relatively difficult to detect any differences in grammar in a group of university students who are generally at a greater level of exposure to printed text through their educational activities. In addition to the challenges in assessing grammar, it is also difficult to measure its development over time for the same reasons. Grammar develops very slowly in a well-educated native speaker of a given language. It is very challenging to select a task that detects grammar development over time as research shows that grammar, unlike vocabulary, develops at a significantly slower rate in adult life than earlier in life (Richards, 1976).

2.4.4 Reading and writing and academic achievement

Reading and writing skills are good predictors of overall academic achievement in English L2 university students. Due to a shortage of reliable instruments to assess reading comprehension and writing in adult English L2 learners, researchers in this area usually explore IELTS-like reading and writing tasks. Another reason for using these tasks is that

these two subscales of IELTS show a good relationship with ultimate academic achievement at university (Bellingham, 1995; Ferguson & White, 1998; Kerstjens & Nery, 2000; Yen & Kuzma, 2009). IELTS-like reading and writing tasks were explored in a study by Harrington & Roche (2014). The authors investigated the importance of vocabulary, reading, and writing in university students in Oman. All of the participants (*N* = 174) studied in the fields of humanities and social science, business, and engineering and information technology.

Academic achievement was conceptualised as GPA at the end of the second semester, constituting the outcome variable. Alongside the IELTS-like reading and writing tasks, the TYN assessment of vocabulary was also administered. In the writing task, participants had to produce a 250-word essay that was subsequently marked by an IELTS professional. The reading task consisted of two 850-word passages followed by 27 comprehension-checking questions. The time limit for each task was 40 minutes. The results showed that writing was the variable that yielded the highest and significant correlation with the GPA. A following regression model including the measures of reading, writing, and vocabulary explained 30% of the variance in students' marks (R^2 = .323), with writing alone being the strongest predictor, explaining over one quarter of the variance.

Reading and writing proved to be good predictors of academic achievement in Chinese L1 students at a UK university. In the study by Trenkic & Warmington (2019), cited on multiple occasions above, both skills were investigated in terms of their importance in overall academic achievement. Reading skills were assessed using *The History of Chocolate*, a 492-word passage followed by 15 comprehension-checking questions, with results measured in terms of the proportion of correct answers provided. Writing was operationalised as summarisation skills, as participants in this study were asked to write a summary of *The History of Chocolate*. The summary was scored in terms of the number of relevant content points recalled in the summary. The results showed that both measures correlated significantly with students' GPA: reading, *r* = .381; writing, *r* = .365. Taken together, both measures (reading and writing) also predicted almost 10% of unique variance in students' GPA.

2.4.5 Skills that underpin learning and language development

The abovementioned aspects of English knowledge and skills are underpinned by cognitive abilities. Intelligence and working memory are among the cognitive abilities that not only underpin learning and second language development, but also directly predict ultimate academic attainment. This section will therefore define the constructs of intelligence and working memory and provide some evidence demonstrating their importance in learning and academic achievement.

Intelligence. Intelligence is a construct difficult to define because of its complexity. The most influential paradigms involve the construct of general intelligence (*g*) and specific cognitive abilities. According to the former, intelligence is a unitary construct, while according to the latter, there exist several components of intelligence, with two main types: fluid and crystallised intelligence (Sternberg & Kaufman, 1998). *Fluid intelligence* (*gf*) is the ability to generate, transform, and manipulate different types of information that are novel and independent of content knowledge. *Crystallised intelligence* (*gc*), on the other hand, reflects a person's general knowledge, vocabulary, and reasoning based on previously acquired information. It is a broad ability to use learned knowledge and experience. The tasks that assess intelligence are usually divided into verbal and non-verbal. In the verbal tasks, participants can be asked to repeat a series of digits, define words, solve arithmetic problems, or provide similarities between two items. In the non-verbal tasks, participants can be asked to indicate which part of a picture is missing, arrange pictures to make a sensible story, or solve picture puzzles. Intelligence develops during adolescence, peaks in late adolescence, stabilises in early adulthood, and declines with age (Chierchia et al., 2019). Therefore, the tasks need to be adjusted to the test taker's age.

Intelligence is a well-established predictor of study success measured in terms of overall academic achievement (Busato et al., 2000; Di Fabio & Palazzeschi, 2009; Farsides & Woodfield, 2003; Gardner & Moran, 2006; Harris, 1940; Laidra et al., 2007; Lounsbury et al., 2003; Neisser et al., 1996; Smrtnik Vitulić, & Prosen, 2012). It has a big impact on overall academic attainment in all stages of education, from primary and secondary school (Di Fabio & Palazzeschi, 2009; Laidra et al., 2007) to higher education studies (Busato et al., 2000; Furnham & Chamorro-Premuzic, 2004; Smrtnik Vitulić, & Prosen, 2012).

Research shows that intelligence, irrespective of the way it is defined or assessed, correlates with academic achievement in university students. For example, Lounsbury et al. (2003) investigated the relationship between general intelligence, personality traits, and course grades in 175 undergraduate students of psychology in the USA over a period of 5 years. The correlation between intelligence and course grades proved to be significant and moderate with $r = .40$, $p < .01$. Intelligence was investigated in another study with 91 British university students (Furnham & Chamorro-Premuzic, 2004). A positive correlation was found between their intelligence and grades in one of the university modules. The participants were administered three tasks: one measure of general intelligence, a visual spatial ability test, and a verbal and spatial ability test. The overall module grade yielded a significant positive correlation with all measures of intelligence and the highest with the visual spatial ability test ($r = .25$, $p < .05$). Both measures of intelligence, fluid and crystallised, showed relationship with grades in 203 university students of primary education and 80 students of social pedagogy in Slovenia (Smrtnik Vitulić, & Prosen, 2012). The non-verbal intelligence test accounted significantly for 4% of the variance in the primary education students and the verbal cognitive abilities task explained 7% of the variance in the grades of the social pedagogy students. Research from all around the world and including native and non-native speakers of English shows that intelligence is an important predictor of academic outcome. Due to the fact that there are multiple predictors of study success and that we can expect little variation in the level of intelligence in students at tertiary level, even the low correlational coefficients obtained in these studies are in fact very meaningful.

Working memory. Working memory is a system of temporary processing and storage of information in the performance of cognitive tasks. It is 'a limited capacity system allowing the temporary storage and manipulation of information necessary for such complex tasks as comprehension, learning and reasoning' (Baddeley, 2000:2). According to the model of working memory developed by Baddeley and Hitch (1974), it consists of four main components: the central executive, phonological loop, visuospatial sketchpad, and episodic buffer. One of those components, the phonological loop, is assumed to be one of the best-developed components of the working memory model and is responsible for 'the temporary maintenance of acoustic or speech-based information' (Atkins & Baddeley, 1998:537). It 'is assumed to hold verbal and acoustic information using a temporary store and an

articulatory rehearsal system' and it is assumed to comprise a temporary phonological store in which 'auditory memory traces decay over a period of a few seconds' (Baddeley, 2000:3). Its main role is the retrieval of sequential information, reflected most clearly in the memory span task with a sequence of items. The phonological loop is best measured by the digit span test involving immediate serial recall of strings of numbers (Atkins & Baddeley, 1998). Working memory, similarly to other cognitive abilities, declines with age, with the threshold for significant decline in working memory placed between ages 60–69 and 70+ (Dobbs & Rule, 1989).

Working memory is important in learning and language development because it impacts linguistic processing in the first and second language. It determines the ability to learn new words in the first language (Daneman & Green, 1986) and reading comprehension in the first language (Daneman & Carpenter, 1980; Waters & Caplan, 1996). Working memory is even more important in vocabulary learning in a second language (Atkins & Baddeley, 1998) and in second language reading comprehension (Chun & Payne, 2004; Harrington & Sawyer, 1992; Lesser, 2007). Reading is a very complex cognitive task that involves decoding the linguistic information from the text, integrating the extracted information into phrases, sentences, and paragraphs, and synthesis of all the information (Koda, 2007). The goal of reading is to construct a meaning based on visually encoded information. All these activities are cognitively demanding, and good working memory capacity is needed to perform these tasks successfully. Reading in a second language is even more demanding because L2 reading involves 'dual language involvement in each operation' (Shahnazari-Dorcheh & Adams, 2014:20), involving more cognitive load and, consequently, greater working memory capacity.

Research shows that working memory is not language-independent and people tend to have higher working memory span results in their first language than in their second language (Osaka et al., 1993). This does not mean that they have lower working memory in another language – they just have lower scores. This is probably because of the properties of subvocal rehearsal in the phonological loop, as it is easier to store or rehearse phonologically entrenched sequences of the L1. Also, processing in the L2 taxes the working memory more than processing in the L1. However, results obtained for the first and second language are usually related. For example, Osaka & Osaka (1992) found very high

correlations between working memory span results obtained for Japanese learners of English. Students who demonstrated high working memory spans in Japanese also tended to have high spans in the English version of the test. In sum, working memory and intelligence are very important in literacy development and researchers should measure both constructs when investigating language and its development in a university context.

This section has shown that language is a very strong predictor of international students' academic achievement defined in terms of abilities to read and write, or grades obtained at university. The importance of language in attainment is sometimes questioned as studies on the predictive validity of IELTS and TOEFL have yielded low correlation coefficients and mixed results in general. However, the reasons behind these results were discussed here and it was shown that small correlations can be meaningful and that there exist many extraneous variables that may distort the picture of findings. Studies that investigate the importance of vocabulary, reading, and writing show much stronger evidence when compared to the predictive validity of the standard proficiency tests. One of the reasons behind this may be that IELTS and TOEFL were not developed to predict academic attainment in the first place and there is a very narrow range of scores, which is problematic for analysis. Vocabulary, reading, and writing may bring more consistent results as the scores in such tests have a greater range. It is also important to bear in mind that different courses demand different skill sets, making the overall test scores less sensitive to detect meaningful results. One more reason may be that the test scores do not reflect true knowledge due to, e.g., test-taking preparation strategies. Answering this question is beyond the scope of this thesis, but the most important conclusion is that general knowledge of English (vocabulary and grammar) and skills in reading and writing are important in learning and tend to be good predictors of outcomes for English L2 students at university.

2.4.6 Summary so far

International EFL students in English-medium universities seem to pursue their studies with a systemic disadvantage, as they enjoy a lower level of success at university when compared to home students. Research shows that international students arrive at universities with a

significantly lower level of English when compared to native speakers of English, their language is not aligned well with the linguistic demands of the courses, and the level of language with which international students begin their studies predicts how much they can achieve academically. Despite a lack of direct evidence, there are also good reasons to think that international students cannot compete with the home students academically until they catch up linguistically. It is therefore possible that the underachievement in relation to home students may occur because the level of English at the start is not sufficient to allow international students to catch up linguistically in a timely manner. More research is needed on the language of international and home students as very little is known about how their language develops at the initial stages at university. There exists only one comprehensive study and more research needs to be done in this area.

Researchers need to adopt a more comprehensive assessment of language and literacy skills as they are much better predictors than the standard proficiency tests such as IELTS and TOEFL, and it is crucial to identify the skills that are especially important. Vocabulary, reading, and writing seem to be very important and need to be investigated further, but very little is yet known of the importance of grammar for academic achievement. In addition, reading and writing need to be investigated using a wider range of tasks other than IELTS-like tasks that seem to be prevalent currently. Only one study cited above (Trenkic & Warmington, 2019) controlled their participants for cognitive abilities and this kind of assessment needs to be adopted more widely. Participants also need to be controlled at least for the subject studied and their first language, as these factors can distort the pattern of findings. Some of the studies cited above took a glimpse at the speed of language processing and spelling, but still very little is known about how important these skills are. A comprehensive task battery would ideally include tasks that tap into cognitive abilities (intelligence and working memory), knowledge of language (vocabulary and grammar), literacy (reading and writing), and phonological skills.

2.5 Chinese vs other international students

As stated in section 2.3.1, international students from China represent the vast majority of all international students in the UK. Research shows that their level of academic attainment

seems to be lower when compared not only to British home students (Crawford & Wang, 2015), but also to other international students (Iannelli & Huang, 2014; Li et al., 2010). At the same time, they struggle linguistically more than other international students (Edwards et al., 2007; Gu & Maley, 2008; Jin & Cortazzi, 1996; O'Connell & Resuli, 2020; Yen & Kuzma, 2009; Zeng, 1996), and their second language skills seem to be the greatest obstacle while at university (Evans & Morrison, 2011; Zhao & Mawhinney, 2015). Chinese is a language very different from English in all aspects of linguistic analysis (this will be discussed in more depth in section 2.6.2), which may be responsible for these students' difficulties with English and consequently their ultimate level of achievement while at university. If second language skills predict academic achievement in English L2 students, it is likely that those speaking first languages that are very different from English may be impacted to a greater extent than speakers of languages that are more closely related to English. This is because of the additional effort Chinese L1 students need to put into mastery of the second language. This in turn may impact academic achievement in this group to the greatest extent.

2.5.1 Chinese L1 students attain less than other international students

Chinese L1 students not only perform less well than British home students in UK universities (Crawford & Wang, 2015; Trenkic & Warmington, 2019), but also underperform in relation to other international students. For example, Iannelli & Huang (2014) explored Higher Education Statistics Agency (HESA) data to investigate the pattern of participation and performance of international undergraduate and postgraduate students in UK universities at three points in time between 1999 and 2009. The study found that Chinese L1 students were less likely to obtain a good degree in comparison to home students and other international students. The odds ratio of obtaining a good degree for Chinese students was only 37% of that for UK students and 52% of that for other EU students, with all differences statistically significant.

In addition, the level of academic attainment in the undergraduate Chinese sample has been decreasing over the time span investigated. Chinese L1 students were most likely to obtain lower second-class degrees, followed by upper second-class degrees. The number of Chinese L1 students who obtained a third-class degree has increased from 14% to 21% and the number of graduates holding an upper second-class degree has decreased from 50% to 43% over the investigated time span. A closer examination of the Chinese sample

demonstrated that those students who obtained their A levels or Highers in the UK prior to the start of their degree were significantly more likely to obtain a good degree than those without such qualifications. This may be because of better developed English skills or experience with the local educational system.

In another study from the UK, Li et al. (2010) also demonstrated that students from China performed less well than other international students. The authors compared Chinese and other international master's students (*N* = 178) enrolled in the School of Management in the academic year 2005−2006. The Chinese L1 students comprised nearly half of the international sample (49%), while the rest of the students came from countries including Greece (10%), Thailand (9%), Nigeria (8%), Taiwan (7%), and Korea (2%); in total, the students represented 25 different nationalities. The results showed that the subsample of Chinese L1 students began the university course with a lower level of English proficiency (as indicated by self-reported proficiency test scores) than other international students, and that they attained lower grades after the first semester.

The self-reported marks attained at the end of the first semester constituted a dependent variable in the investigation of the predictors of study success. The following potential predictive factors were included: writing ability, perceived value and importance of learning success to family, learning preference, effort, familiarity with environment, and socialisation with compatriots and others. The results showed that it was writing ability that predicted the achievement differences between the Chinese L1 and other international students. The Chinese L1 students' writing ability was significantly lower when compared to that of other international students, and writing ability proved to be the second most important predictor of all international students' performance, ranking just below the perceived value of learning success to family.

2.5.2 Chinese L1 students struggle linguistically more than other international students

In addition to the attainment differences, research shows that students from China demonstrate weaker skills in English when compared to other international students. In accordance with Li et al. (2010), cited above, skill in writing has been confirmed as an important skill differentiating Chinese L1 students from other international students. This

was demonstrated in a study on an IELTS research programme by Mayor (2006), who focused on Chinese and Greek L1 students' IELTS written work. Participants were asked to write an argumentative essay; they were instructed to write a minimum of 250 words and were given 40 minutes to complete the task. The data comprised a corpus of 186 essays, and the analysis focused on specific discourse features such as the use of personal pronouns (*I* and *you*) and the choice of verb mood (interrogative and imperative clauses). The results showed that Chinese L1 students used a greater number and wider range of interpersonal pronouns and used a greater proportion of interrogatives and imperatives when compared to the Greek L1 students and established academic practices. The authors concluded that these features may lead to poorer IELTS scores and that this practice may stem from the IELTS test preparation practices that are popular especially in China.

In another study, Trice (2003) investigated the perceptions of academic staff about international students in four departments at a US university (architecture, public health, mechanical engineering, and material science and engineering). In their interviews, faculty members mentioned language difficulties more often than any other issues across all four departments. They also commented on the differences within the whole international group and that 'significant variety exists among international students based on their country of origin' (p. 390). Due to the fact that the vast majority of students across all departments came from Asia, and specifically from China, professors had these students in mind when reflecting on language issues. They claimed that '[s]tudents from Europe often arrive with a better command of English and develop relationships with Americans more easily' (p. 390). Academics' perceptions were also utilised in a study by Yen & Kuzma (2009), who found that international students speaking Chinese L1s seem to be struggling linguistically at universities to a greater extent than other groups of international students. The academics from one of the universities in the UK stated that language problems are concerning, especially in relation to students speaking Chinese L1s. This is supported by evidence from interviews with academics who commented on Chinese L1 students relying on personal interpreters in their lectures and having difficulties with written assignments, argumentation, and understanding assignment criteria.

The existence of regional variation in terms of language-related adjustment to English-medium universities has been confirmed in another study on adjustment to university in the US context (Senyshyn et al., 2000). The authors specifically investigated 'country clusters' based on participants from south and east Asia (54%), western Europe and Canada (23%), central and east Europe (13%), and central and south America (10%). The questionnaire administered to students consisted of two main parts: satisfaction and confidence in adjustment and difficulty with adjustment. There were a number of questions on reading and writing experiences, participating in class discussions, and conversations in English. According to the results, nationality was an important variable in the process of adjustment. The participants from western Europe and Canada adjusted better and had fewer problems than the Asian students.

2.5.3 Chinese students' language-related experience

English language related problems are well documented and the most common issues encountered by Chinese L1 students while in English-medium universities. In a study on the academic challenges encountered by Chinese engineering transfer students in a US university, 44 participants responded to a four-part online survey (O'Connell & Resuli, 2020). The study found that the two most serious academic issues experienced by the international students were difficulties in obtaining appropriate transfer credit and difficulties in understanding their instructors' English. The latter was indicated by 73% of the participants taking the survey. What is more, language problems impacted studies indirectly as it was found that language issues had a negative impact on other activities such as asking questions and discussions in class, group work, and forming study groups. These difficulties were exemplified by rich qualitative data, with participants stating overtly that their English skills were not sufficient to cope with the demands of their course. The participants were also given the option to comment on the most challenging academic issues. These data confirmed issues with English as the main source of difficulties while at university and these included lack of oral skills, lack of understanding of technical concepts and vocabulary, and problems with reading in English.

Further evidence suggests that what is perceived as culture-related adjustment to English-medium universities in fact stems from lack of confidence in English. Studies suggest that issues that are sometimes ascribed to cultural differences, or even culture shock, may be a consequence of low proficiency in English leading to a lack of confidence in speaking and in taking an active part in seminars. This is confirmed by Gu & Maley (2008), who were interested in the influence of language, among other factors, on adaptation to academic life in Chinese students from a variety of courses – foundation, undergraduate, and postgraduate – in a UK context. They obtained their data from questionnaires administered to 163 students in four universities and colleges in the UK, and interviews with 28 undergraduate and postgraduate students from 10 universities. Their findings suggest that the language barrier may be responsible for issues in learning and communicating in English.

In another study, Jin & Cortazzi (2006) also found that Chinese L1 students struggled with communicating spontaneously in seminars and group work without preparation, which was linked to lack of practice and confidence in oral English. Some difficulties, however, are purely culture-bound. For example, Chinese L1 students struggled with essay writing and this problem was attributed to lack of training in their home country and lack of knowledge of the discourse patterns expected in the UK. The Chinese L1 students were unfamiliar with expectations for written assessment, criticality, and expressing opinions. There were also problems with acknowledging sources and referencing in writing, which could originate in a Chinese education system. This shows that cultural differences indeed exist, but they can be exacerbated by the level of English that Chinese students begin university with, which may not allow them to overcome these problems in a timely and effortless manner.

To further highlight the linguistic factor in adjustment to English-medium universities, it is worth examining some evidence from a context where cultural differences and culture shock are not expected. For example, in a study by Evans & Morrison (2011), the authors tracked 28 participants throughout their undergraduate degree with English as the language of instruction at a polytechnic in Hong Kong, where 90% of the population use Cantonese as their first language. The participants studied within a number of different programmes and were interviewed throughout the three-year programme. Chinese students identified problems with English as the most important issue that affected their academic

achievement. Specifically, they had difficulties in understanding technical vocabulary, comprehending lectures, and achieving an appropriate writing style. In addition to the interview, all the participants were administered a 45-item survey about their academic English skills at three points in time during the three years of the study, where they assessed their degree of difficulty with the four language skills.

The skill that proved to be the most difficult was writing (using appropriate academic style, expressing ideas in correct English). The problems related to reading included understanding specialist vocabulary and working out the meaning of difficult words. The problems with speaking included use of correct grammar and pronunciation. Reading-related problems in Chinese L1 students appear elsewhere in the literature: for example, Zhao & Mawhinney (2015) found in their qualitative study that Chinese L1 students in Canada have slower reading speeds when compared to native speakers of English and that one of the reasons for this is lack of knowledge of technical and specialist vocabulary.

It has been shown that Chinese L1 students attain less academically than home and other international students, struggle linguistically more than other international students and their adjustment to universities is driven mainly by the problems with English. Since language is at heart of what international students can attain academically, it is therefore likely that Chinese L1 students under attainment may have its origins in their skills in English. The differences in performance can be also attributed to cross-cultural differences, and it is claimed that Chinese L1 students are reluctant to ask questions and participate in class activities because of their cultural background. They must adjust to different rhetorical style, writing and referencing conventions, values, and educational system, and culture shock is sometimes mentioned as the reason behind the attainment differences. However, Chinese L1 students can perform similarly to other students after the first year at university (Crawford & Wang, 2015); it was found that those Chinese students without experience of studying abroad performed significantly better than their compatriots with such experience (Li et al., 2010), and that academic achievement is at risk in a context where culture shock is not expected (Evans & Morrisons 2011).

Research shows that it is sometimes lack of oral language skills that stops students from participating in seminars, exchanging ideas, understanding assessment criteria, and seeking help from the teaching and support staff. The evidence on the actual level of English

in Chinese L1 university students is very limited and only one study so far has investigated language skills in Chinese L1 students in a comprehensive way. To the best of my knowledge, no study to date has compared international students' language when controlling for the first language spoken or cultural background. Therefore, more research is needed to better understand language and literacy in different cohorts of international students and their importance for academic achievement.

2.6 Cross-linguistic differences between English and Chinese

2.6.1 The notion of language family and linguistic transfer

Research in the field of second language acquisition demonstrates that individual learner differences such as the first language spoken may affect the rate and trajectory of language development (Lightbown & Spada, 2013). Different combinations of L1s and L2s can influence language learning in different ways due to linguistic transfer. Linguistic transfer can be defined as the use of linguistic (and cognitive) knowledge acquired in the L1 for L2 learning (Odlin, 1989), and it hinges on the linguistic distance between the L1 and L2 (Koda, 2007). An example can be given from research on reading: 'When the L1 and L2 are closely related, shared structural properties pose similar demands on processing and allow L1 competencies to function in L2 reading with little adjustment. By contrast, L1 skills do not facilitate L2 reading to the same extent when the two languages are distantly related' (Pasquarella et al., 2015:1). This also applies in research on vocabulary learning, as novel words consistent with native language phonology have been shown to be learned faster than words with unfamiliar sounds, suggesting that prior knowledge impacts the process of novel knowledge integration (Havas et al., 2018). In general, the L1 can facilitate processes such as reading and vocabulary learning in a second language through the impact of characteristics shared between the L1 and L2.

Linguistic proximity between languages is explored in this study through the notion of language family. A language family is a group of languages derived from a common ancestor and sharing 'certain observable linguistic characteristics, such as words, sounds and grammatical patterns' (Pereltsvaig, 2012: 8). The similarities across a single language family are established using comparative reconstruction methodology by compiling lists of cognate

words. The cognate words include numerals and other basic vocabulary such as words for body parts, kinship relations, and natural phenomena – words that are not readily borrowed from other languages and are therefore assumed to originate in a common ancestor language. In addition to lexical items, comparisons are also made in relation to sounds and to grammatical patterns such as word order and case forms. Languages that share a common ancestor are grouped in so called 'language families'. The languages of the world are grouped into nine main language families, further divided into smaller branches. English and Chinese belong to two distinct language families: Indo-European and Sino-Tibetan, respectively. The Sino-Tibetan branch is divided for smaller branches with Sinitic language family to which Chinese language belongs. This typological distance makes these two languages very different from each other. In contrast, English, and other languages within the Indo-European family of languages, to which almost all European languages belong, are more closely related and more similar to each other. Therefore, L1 speakers of Chinese and L1 speakers of most European languages may differ in their experience and speed of learning English (Pereltsvaig, 2012).

The next section describes the main linguistic features that differentiate English and Chinese and that may affect learning speed and lead to some potential difficulties for Chinese L1 speakers while at university.

2.6.2 English vs Chinese

English and Chinese belong to two very distant language families, Indo-European and Sinitic, respectively. This section describes the main linguistic features that differentiate these two languages to shed light on the additional workload involved in the experience of learning English for L1 Chinese speakers. It will outline the main features, including the writing system, vocabulary, phonology, grammar, and morphosyntax.

Writing system. The first and most striking difference between Chinese and English is the writing system. English uses an alphabetic writing system, which is a script based on individual symbols or groups of symbols that represent phonemes, the smallest units of speech. There exists a very close correspondence between the symbol (grapheme) and the phoneme. Alphabetic orthographies can be divided into shallow and deep. In shallow orthographies, the correspondence between grapheme and phoneme is very consistent,

and this is true for languages such as Greek, German, Finnish, Spanish, and Italian. In deep orthographies, on the other hand, as for example in English and Hebrew, the correspondence between grapheme and phoneme is not so regular: a single grapheme can represent several different phonemes (e.g., the grapheme *g* in English can represent different sounds in different words, e.g., *ginger*, *guilty*), and a phoneme can be represented by different graphemes (e.g. phoneme /k/ in English can be represented by two different graphemes: *c* and *k,* e.g., *cake, kitten*) (Ellis et al., 2004).

Chinese, on the other hand, is a language using logographic writing system where a character is the basic unit of writing. Most of the characters are compound characters consisting of radicals, with two main types: the semantic and the phonological radical. The semantic radical conveys information about the meaning of the word: it can indicate the conceptual category of the character (e.g., a radical *female* in the word for *mother*) or be directly related to the meaning of the character (e.g., a radical *wood* in the word for *cabinet*). However, there is a high portion of semi-transparent characters where the semantic radical does not contribute to the meaning of the word directly (e.g., the animal radical in the word for *hunting*). There is also a handful of characters in which the semantic radicals do not provide information about the character's meaning. The second radical type is the phonological one, and it gives an indication of the character's pronunciation, but again, the pronunciation rules are not always regular. The phonological radical can give full information on pronunciation (onset, rime, and tone), partial information (onset and rime), or no information. The internal structure of a character (e.g., the position of a stroke in the character) and the position of the components (semantic and phonological) are important in character recognition. These features make the whole writing system very complex visually, as each radical, the basic unit of writing, consists of several individual strokes (Ross & Ma, 2009).

The main difference between languages using an alphabetic writing system and those using a logographic writing system, such as Chinese, is that the second type are not regular and transparent in terms of the orthographic-phonological relationship. This may encourage different reading processes; for example, each type of writing system may involve a different set of skills for script recognition. According to the orthography depth hypothesis, readers of shallow orthography languages, such as Spanish, use phonological information to

a greater degree than the readers of deep orthographies, such as English or Hebrew (Ellis et al., 2004). Furthermore, the Chinese writing system is far more complex visually when compared to the alphabetic system. It is now well established in the literature that reading alphabetic script involves phonological skills, whereas reading logographic script involves good visual and orthographic skills. This difference puts Chinese L1 speakers at a disadvantage while reading in English when compared to those speaking European L1s, whose phonological skills are better developed because of their familiarity with alphabetic writing scripts.

Vocabulary. Another challenge that Chinese learners of English must face when learning English vocabulary, and which is not faced by learners who speak European L1s, is lack of cognates. Cognates are words shared across two or more languages that have similar meaning, spelling, and pronunciation, and found in languages sharing a common ancestor. It has been demonstrated that cognate awareness can facilitate novel word learning, and since English and European languages share cognates, this gives European L1 speakers an advantage in English vocabulary acquisition. In addition to the presence of cognates, many vocabulary items in English and European languages are related semantically. Some of the words in European languages may have the same root that has diverged in meaning over time, e.g., English *loaf,* Russian *xleb,* and Polish *chleb* meaning *bread*. On the surface, these words seem rather different. However, the English word derives from *hlaf*, meaning 'bread, loaf of bread'. This word has lost its initial sound 'h' and its meaning has narrowed to denote a unit of bread, and not the substance as a whole (Perelstvaig 2012: 17). This semantic relatedness among vocabulary items in languages derived from the Indo-European language family can be facilitating for English learners speaking European L1s.

It has also been demonstrated that learning words written in an alphabetic writing system poses great challenges for Chinese L1 users. Chinese characters are very different from a string of letters making a word in an alphabetic writing system. In logographic script, perception is focused on a number of interrelated lines making up the character, occupying a fixed space on the surface, whereas alphabetic words are strings of letters of varying length. Take the two words *conversation* and *conservation* – distinguishing them involves a great deal of perceptual discrimination. In Chinese, a character corresponds to one syllable and words consists of one or two characters; therefore, a long string of alphabetic words

may pose great difficulty for pronunciation and spelling. Compare a one-syllable Chinese word and an English word such as *categorisation*. When it comes to word form, most Chinese nouns consist of two characters, each of them being a related concept, and when put together, they are given a new meaning: for example, the word for *plane* consists of two words: *flying* + *machine*; the word strike is built from *stop* + *work*. Therefore, the concepts involved in English word creation may be challenging to grasp, and problems at the semantic level may arise (Ma & Kelly, 2009).

Phonology. According to the taxonomy of natural languages, they can be divided into two main groups in relation to accent: accentual and non-accentual. Mandarin Chinese is a non-accentual tone language, whereas English and other Indo-European languages are accentual stress languages. Chinese is a tonal language, which means that it distinguishes different tones that can determine the meaning. There are four tones in Chinese: (1) a high even tone, (2) a rising tone, (3) a falling-rising tone (called dip tone), and (4) a falling tone. Tone may distinguish not only lexical but also grammatical meaning. The basic phonological unit is a syllable broken down to onset and rime. In Mandarin Chinese, the onset is always a single consonant (initial clusters are not allowed), and the rime consists mainly of a vowel. The notion of tone is also present in English and other European languages, but it is used for intonation, for example to express emphasis, to convey surprise or irony, or to distinguish a question from a declarative sentence. Furthermore, in non-tonal languages such as English, pitch or tone is a property of utterances, and not of single words (Ross & Ma, 2009).

Morphology and syntax. In English and most European languages, the word creation process is based on derivational morphology. For example, the suffix *-ize* is added to English adjectives and nouns in order to create verbs, and *-y* turns a noun into an adjective. A similar operation works, for example, in Spanish, where suffix *-oso* converts nouns into adjectives and *-cion* changes verbs to nouns. Chinese, on the other hand, uses compound morphology in creating different word classes. Over 70% of Chinese words are compound words formed by a combination of two roots, and the number of derivational morphemes in Chinese is smaller than in Indo-European languages (Li & Thompson, 2003). One of the greatest grammatical differences distinguishing Chinese and English (and by extension many European languages) is that there is no verb inflection that can express grammatical notions such as tense or aspect. In Chinese these features are represented by word order and the

use of grammatical particles. Chinese also demonstrates a distinct word order, has no determiners, and marks its plural form in a completely different way in comparison to English.

These cross-linguistic differences show the additional burden involved in learning English for those who speak Chinese L1s. These large differences are likely to impede linguistic transfer and learning English becomes slower and more challenging.

2.6.3 Summary so far

International students from China are especially important in UK universities because they are the most numerous across universities in the UK. They are important in economic terms and also from a pedagogical point of view, through enriching the learning environment for local home students. The previous sections have established that the level of language with which international students arrive at university is critical in academic achievement, specifically in terms of knowledge of vocabulary, reading and writing. We also know that Chinese students are very distinct culturally and that they usually arrive with lower skills in English when compared to other international students and attain less academically. Therefore, more research is needed to understand the language skills in this group and their importance in academic achievement. It is also important to find out if the findings obtained in Trenkic & Warmington (2019) generalise to students who are linguistically and culturally closer to English. This study aims to fill this gap as it compares language and literacy skills in British home students and in international students who speak first languages with different degrees of relatedness to English and who are distinct culturally. This is done to determine whether the previous findings generalise to different groups of international students.

2.7 Summary and research questions

This literature review has shown that proficiency in the language of instruction is critical in terms of international students' attainment. They constitute one of the most important factors that impact academic attainment at university. The skills that have proved to be especially important are reading, writing, and knowledge of vocabulary. Phonological skills have also emerged as important in underpinning language and literacy development. It is

therefore crucial to investigate these skills in British home students who study in their first language and international students pursuing higher education in a second language. The group of international students is itself heterogeneous in terms of the first language spoken and it is crucial to understand the differences among them, as research in this area is lacking. More research is needed to compare their actual skills in English using more comprehensive methods of language assessment and controlling for the first language spoken. This study is designed to fill these gaps and extend the findings obtained in Trenkic & Warmingtom (2018) to find out if the differences found between home and international students extend to different cohorts of international students. This study therefore aims to answer the following two research questions:

RQ1. How much do knowledge of the English language and literacy skills differ at the start of the first year at university between:

a) Home students who speak English as their first language and international students who speak English as a foreign language (EFL)?
b) Those speaking European L1s and Chinese L1s?

RQ2. How much do knowledge of English and literacy skills change in the first year at university in home students and in international students speaking Chinese and European L1s? Do international students close the gap on any measures?

In relation to the first research question, it is predicted that the group of British home students will have stronger language and literacy skills when compared to the group of international students at the start of their course. This prediction is made on the grounds of the first language spoken and the existing literature showing that native speakers of English have stronger skills when compared to English learners. This was found in studies examining vocabulary, reading, and writing (Devos, 2019; Elder et al., 2007; Morris & Cobb, 2004), as well as reading comprehension and phonology skills (Pasquarella et al., 2012; Trenkic & Warmington, 2019). To date, research found large differences in language and literacy skills between home and international students, but the international students in those studies were exclusively or predominantly Chinese (Trenkic & Warmington, 2019), which raises questions of population validity. In particular, the findings may not generalize to students who speak languages that are typologically closer to English, or who have had more

exposure to English in daily life, either through geographical closeness (ease of travelling) or cultural closeness.

With respect to differences between EFL with Chinese L1s and EFL with European L1s, it is predicted that students speaking European L1s will have stronger language skills. This hypothesis is based on their first language status and the existing literature suggesting that students speaking Chinese L1s may have less well-developed skills in English when compared to other international students at the start of university. However, the actual skills in English had never been compared directly in these two groups of students and this prediction needs to be treated with caution.

It is predicted that all three groups of participants (ENS, EFL with European L1s, and EFL with Chinese L1s) will improve their linguistic profiles in most of the measures across one year. This is expected as a natural process of language development in the second language country for the EFL students, and of language development resulting from a high volume of exposure to academic language and materials in the group of ENS. When it comes to the level of improvement, different predictions could be made on different grounds. Logically, the group that demonstrates the lowest skill could improve the most because of the greatest scope for improvement. At the same time, the group that demonstrates the lowest skills at the start could improve the least because of lack of abilities and motivation for learning when material is too difficult. There may be other factors that come into play at the time when students arrive at university, such as the level of exposure to English. Therefore, the question of which group would be expected to improve most may be an open question rather than one where a firm prediction of the direction can be made.

In the next chapter, I present the methodology, describing the rationale behind the study design. The chapter describes participant selection and the recruitment process, the rationale for selecting the battery of tasks, and the statistical methods used to analyse data.

Chapter 3: Methodology and methods

The present study builds on the findings presented in the literature review and aims to fill the research gaps identified. It is a longitudinal investigation into language and literacy skills in English native speakers (ENS) and English foreign language learners (EFL) studying at a UK university. The EFL students were selected according to their first language, with two main groups: EFL with European L1s and EFL with Chinese L1s. The students referred to as Chinese L1s consist of L1 speakers of Mandarin and Cantonese, two main dialects spoken in China and belonging to the Sinitic branch of the Sino-Tibetan language family, which is very distant from the Indo-European family to which English belongs. The EFL with European L1s speak languages that belong to the same language family as English. This is at the heart of the design of the present study, as it investigates whether the previous findings based on students speaking Chinese L1s can be generalized to other groups of international students. All three groups of participants were administered a task battery tapping into linguistic knowledge (vocabulary, grammar), literacy skills (reading and writing), and also phonological skills, at two time points: at the start of their undergraduate university course in year 1 and one year later, at the start of the second year. The two points of data collection are referred to as Time 1 (T1) and Time 2 (T2).

This study involves an approximate replication of Trenkic & Warmington (2019). Replicating previous studies as a serious research methodology has emerged in applied linguistics relatively recently. It embraces a series of modified repetitions of the original experiment along a continuum and offers not only validation of previous findings, but also useful contributions to the field (Porte & McManus, 2019). There are several differences between the original and the present study. Firstly, the choice of participants in the original study was restricted by the learning context where it was not possible to compare students at the same level of study. The present study overcomes this limitation by comparing students at the same level as they all are commencing their undergraduate courses. This allowed a longer span between the two testing sessions and unlike in the original study, where all the participants were re-tested after seven months, in this study all the participants were re-tested after one calendar year, at the beginning of their Year two at university. Another modification to the original design was inclusion of one additional group of international

students to find out whether the findings obtained from Chinese L1 students can generalise to other groups of international students. Finally, in terms of the tasks in the battery administered at both time points, the present study used only one, receptive vocabulary test, unlike the two vocabulary tasks in the original study. Reading comprehension and accuracy tasks in this study were assessed with different instruments. The present study also included some additional measures of grammatical knowledge, working memory and writing speed. Unlike in the original study, all language and literacy tasks were administered at both time points and the T2 data were collected online. In sum, the approximate replication design in this study involved modifications to the methodology, better control over participants' background, different mode of data collection, and the type of statistical analyses that followed the modified design.

This longitudinal study was originally designed to use the same participants, instruments, mode of data collection, and procedure at both time points. This, however, became impossible because of the outbreak of Covid-19. This started at the time when the data collection at T1 was completed and meant that the T2 data collection and task administration procedures had to be modified. Therefore, this chapter describes the original design that was implemented at T1 only, as well as statistical methods for analyzing data obtained at both time points. The aspects that have been adapted to suit the alternative mode of data collection are described separately in chapter 5 before presenting T2 results.

3.1 Study design

3.1.1 Between-group design

The first goal of the present study was to compare language and literacy skills in students speaking English from birth and English language learners speaking first languages that differ in the degree of relatedness to English. This resulted in between-group comparisons of the following three groups: native speakers of English, EFL students speaking one of the European L1s, and EFL students speaking Chinese L1s. Each of the participants was recruited to one of these three groups based on the naturally occurring phenomenon of their first language. Because of this pre-existing difference in participant selection (L1), which at the same time constitutes an independent variable, a wholly random allocation to all three

groups was not possible. The implications of this aspect of the design will be discussed in more detail in section 3.1.4 where threats to internal validity are listed.

This study investigated group level trends to see if the results would align with the findings in the literature on differences in academic achievement. As shown in the literature review, international students (comprising mostly of English learners), tend to obtain fewer good degrees than native speakers of English. In addition, those speaking Chinese as their first language seem to be disadvantaged the most among the whole international cohort regarding their educational outcome. It is therefore important to investigate between-group trends in language and literacy skills to see whether they align with the strand of research on academic achievement. Given the importance of language in academic achievement, it is predicted that literacy is stronger in highly achieving groups and weaker in groups obtaining fewer good degrees. The finding of the present study would then shed some light on the nature of the relationship between language skills and academic outcome. The present study is the first one and much needed investigation looking at students of distinct language background, nevertheless, it would be interesting to see comparisons at individual level as well.

3.1.2 Longitudinal vs cross-sectional study design

Another goal of this study was to track the language development of university students over a time span of one academic year. The dimension of time in linguistic studies can be approached from two perspectives, direct and indirect, resulting in two types of studies: longitudinal and cross-sectional. The present study used a longitudinal design, which involves the same participants taking part in a study at at least two points in time. Therefore, the same participants took part in the study at the start of their first year at university and returned one year later. A cross-sectional study, on the other hand, is an indirect, synchronic comparison of a linguistic phenomenon or language change in two or more samples at different stages of development. This involves data collection from different groups of students at different stages of language development at a single point in time, and inferences are based on the indirect impact of the passage of time on students'

language. This is because a cross-sectional design 'simulates tracking the development of those skills in a single learner over the course of real time' (Podesva & Sharma, 2014:498).

The present study used a longitudinal rather than a cross-sectional design for the following reasons. Firstly, the longitudinal design has advantages over the cross-sectional design as it is based on the passage of real time and, thanks to that, it is very sensitive to complex patterns of individual and group change, as well as gains made over time. Because data come from the same participants, a longitudinal study enables the researcher to 'construct more complicated behavioral models than purely cross-sectional or time-series data' (Ruspini, 2002:26). Secondly, due to their more demanding procedures, longitudinal studies are less frequent and therefore called for by researchers (Melby-Lervåg & Lervåg, 2014; Ortega & Iberri-Shea, 2005). The time span between the two waves of data collection was dictated by one of the research questions, namely, whether the initial differences (if any) persist or disappear over the course of one academic year. Language change in the first year at university is of interest here, as the first year of an undergraduate programme is a transition period for students that prepares them for the second-year examinations that count towards their degree classification and sets them on course for their ultimate achievement at the end of their undergraduate course. Therefore, this longitudinal study involved data collection at two points in time, with the first wave of data collected from three groups of first-year undergraduate students at the start of Year one at university, and the second one at the start of Year two.

3.1.3 Lab-based vs online data collection

According to the original design, this study was planned to be carried out in laboratory conditions that involved face-to-face interactions with participants in group and individual testing sessions. A task battery tapping into the knowledge of language, English literacy skills, phonological skills, and cognitive abilities was administered in controlled lab conditions, which is typical in behavioral and perceptual studies like the present one. The same procedure was originally planned for the second wave of data collection at the beginning of Year two. However, lab-based data collection at T2 became impossible due to restrictions on social contact caused by the Covid-19 pandemic and implemented by the UK

government at the time when the T1 data collection process was completed. It was therefore crucial to find an opportunity to repeat testing at T2 while maintaining social distancing and protecting the participants. One possible way of replacing the conventional face-to-face meetings was to switch to online conference technology as an alternative data collection method. This was the alternative chosen. As a result, all the research instruments administered in face-to-face T1 testing sessions were adapted into a web-based mode of data collection and the T2 data were collected in online meetings using Zoom (version 5.2) conferencing software.

Zoom is one of many videoconferencing software types (see Lobe et al., 2020, for an overview) which support real-time audio and full motion video. It has many useful features that enable the user to mimic a real meeting with interactions and presentation of stimuli. These features include screen share and chat functions, among others. Zoom was selected for the present study in the first instance because it was the university's main option for online meeting software. It was expected that all research participants had access to and familiarity with it because it was the primary tool supporting their distance learning since the outbreak of Covid-19. In addition, Zoom has been used successfully in many qualitative, interview-based, and focus group studies (see Chia et al., 2020, for an overview). It has proved to be a reliable research instrument that can successfully replace a face-to-face interaction: for example, in a study by Matthews et al. (2018), the researchers tested many conferencing software types and Zoom proved to be the most reliable because of good sound quality. Furthermore, what makes Zoom unique among other software types is its safety of use. It stores the information securely and manages user and call metadata, which makes it a secure tool for data collection.

Despite its many advantages, Zoom has some limitations as well. Some researchers have reported technical issues with the equipment experienced by research participants (Archibald et al., 2019); however, this mainly happened in populations not familiar with the latest technology. Participants in the present study were expected to be up to date with the latest hard drive, technology, cameras, and microphones. At the time of the study, they had had at least 6 months of experience using Zoom in distance learning, and it was predicted that they were digitally literate and at similar levels of familiarity with the software. Some challenges, however, were still anticipated, and, to mitigate any potential problems, some

extra time was allocated for each meeting in case of technical issues, problems with connection, or a need for extra guidance on Zoom functionality. Another issue reported by researchers collecting their data on Zoom is a lack of control of participants' environment and disruption in the background. For example, Daniels et al. (2019) reported family members and others disturbing the meetings. As participants in the present study were expected to be in their home environment during the testing sessions, they were asked to stay in a quiet space in both meetings in the invitation to the study.

3.1.4 Threats to internal validity

There were two aspects of the design that could have posed challenges in terms of the validity and reliability of the findings. These were lack of true random group allocation during recruitment and different modes of data collection at the two time points. The first limitation is related to some systematic between-group differences. As mentioned above (section 3.1.1), a lack of true random group allocation may lead to limitations in control over extraneous variables and the observed effect may not be wholly attributable to the independent variable (Field & Hole, 2002). It therefore would make it difficult to attribute the initial differences and language change wholly to the L1 and linguistic distance from English. In the case of the participants speaking Chinese L1, the first language variable is confounded with students' nationality. There may be some national or cultural (rather than linguistic) characteristics that have an impact on university students' language; these may relate, for example, to methods of language instruction and exam preparation strategies. It was also stated in the literature review that China as a country sends the largest number of students abroad and this nationality makes up the greatest proportion of international students in UK universities. This may in turn impact the pattern of language use in this group of students, as they may be more likely to communicate in their first language. In sum, lack of true random allocation in all three samples may limit the possible impact of the independent variable (L1) on the results obtained.

The second limitation is the different mode of data collection at the two time points. Since the two waves of data were collected in different conditions (lab vs online), the first challenge is connected to the extent to which Zoom is able to mimic a real face-to-face interaction and, consequently, the reliability of the comparison between the two sets of

results. The main differences between the two modes of data collection were different levels of control over participants' environments and possible problems with connectivity involving sound delay, which was crucial in audio- and time-sensitive tasks involving participants reading words and digits in timed tasks, for example. Another task that could have been affected by delivering it online was the reading task, as research shows that different aspects of reading (for example, speed and comprehension) can be affected depending on whether a text is read from paper or from a screen (Kong et al., 2018). Nevertheless, it was still possible to analyze the two sets of data and make valid claims about the between-group differences at T2 because the different modes of data collection affected all three groups to the same extent. These two aspects of the design will be addressed again in the discussion chapter.

3.1.5 Ethical considerations

The study was approved by the Ethics Committee at the Department of Education, University of York. All the participants took part in the study voluntarily. They were presented with information about the study prior to signing up for the testing session and signed a letter of informed consent before taking part (see Appendix C). The consent form outlined the purpose of the study, the requirements of participation, and the reward. All participants were offered £30 for taking part across the two time points, with £15 paid upon completing all tasks each year. They consented to participate at both time points and gave permission for their course marks to be accessed. Due to the change in circumstances caused by the Covid-19 pandemic and the switch to a web-based data collection mode, the original consent form was amended and presented to all the participants again before the first session at T2 (see Appendix D).

All participants were informed that the study would be held via Zoom, and that they would have to share their screen in both meetings. The informed consent form was presented before the start of the first task in Qualtrics, an online surveying software. Each of the participants typed their name at the bottom of the online form to consent. The data at both time points were collected and stored following data protection guidelines; participants' data were pseudo-anonymized up to the point of data analysis. The document linking the participants' IDs with participants' names was stored on a password-protected

computer. All the participants were guaranteed the right to withdraw from the study at any stage of the experiment and they had the right to withdraw their data up to 20 days after their last participation.

3.2 Participants

3.2.1 Overall characteristics

All the participants were recruited from among students enrolled within the same HE institution. All students interested in taking part in the present study were presented with a screening survey first to check their eligibility (see Appendix B). It consisted of questions on their study programme, year and mode of study, first language, and language-related disabilities. Only those who matched the selection criteria were allowed to sign in for the testing sessions. All the participants taking part in this study were therefore pursuing their first degree and none of them had studied at a tertiary level before commencing their studies. This guaranteed that all the participants were unfamiliar with the university system, grading, and types of assignments. Having studied at university before would be seen as an advantage in comparison to someone without such experience. They were full-time students planning to stay at the university in question for the duration of the whole degree programme. This means that none of them was an exchange student, which would have made it impossible for them to return for the second testing session scheduled at the beginning of their second year.

In participant recruitment, I looked to recruit students from disciplines that are linguistically demanding and to avoid recruiting students from disciplines that are strongly numerically based. Previous research suggests that language plays a more critical role in the former (Hu & Trenkic, 2021) and that aggregated analyses could statistically obscure or even reverse patterns of results. In the end, the final sample of 177 participants comprised 135 students from the humanities and social sciences and 42 from the faculty of science. The latter included students studying biology, chemistry, or computer science, but no participant was studying mathematics, engineering, accounting, or physics, disciplines with stronger numerical focus. None of the participants was diagnosed with dyslexia or any other disability that could affect language, reading, writing, or hearing.

A close examination of participants' backgrounds as reported in an additional background questionnaire led to the exclusion of certain student profiles. This was the case for 6 ENS students who, despite being native speakers of English, were new arrivals in the UK, meaning that they had obtained their education outside of the UK. These students were excluded to maintain the homogeneity of educational experience within the ENS sample. All the EFL students were new arrivals in the UK at the time of the study and all the EFL with European L1s came from European countries. None of the EFL students had grown up bilingually as indicated by their parents' first language. Several were spoken to in English in childhood, but their vocabulary profiles suggested that this did not have a big effect on their proficiency in English as their scores were within the range of their language group. A comparable number of students in each EFL group (29 in the EFL with European L1s and 27 in the EFL with Chinese L1) had experienced education with English as the language of instruction. This ranged from a couple of months as an exchange student, through periods in primary and secondary school, to GCSEs and A Levels. The variety of this experience made it impossible to compare the two groups of students in this dimension in a meaningful and reliable way. Twenty-two EFL students with European L1s and 16 with Chinese L1s had lived in an English-speaking country prior to the start of their degree. This experience ranged between 3 months and 10 years in the former and 4 months and 6 years in the latter sample. Based on these data, it can be concluded that the group of ENS students was homogeneous in terms of their educational experience and the two EFL groups were comparable in terms of their English learning experience.

The whole group, irrespective of their first language, proved to be homogeneous in terms of their socio-economic status as well (see Figure 3.1, error bars represent standard deviations). Socio-economic status was operationalized in terms of caregiver education level. Each participant was asked about the level of education of each parent or primary caregivers. They had to select an option that was the closest to parental education level from a list of six possible answers presented to them in the background questionnaire. The items in the list were placed on a scale from the lowest to the highest qualification level, as follows: 1/ some or no secondary education, 2/ secondary school education, 3/ post-secondary education with vocational training, 4/ university degree, 5/ post-graduate degree or professional education, 6/ don't know/not applicable. Answers 1 through 5 were

allocated the corresponding numbers, which were used to calculate the mean scores for each parent in each language group. There were only a few cases where a student was not able to answer this question and these answers were excluded from the analysis. The mean scores showed very little variation, with the range between 3.5 and 3.8. On average, both parents in all three groups obtained post-secondary education to university degree. The final sample that contributed data to the analysis (*N* = 177) consisted of 123 females and 54 males; the mean age was 18.67 (*SD* = .92) years old.

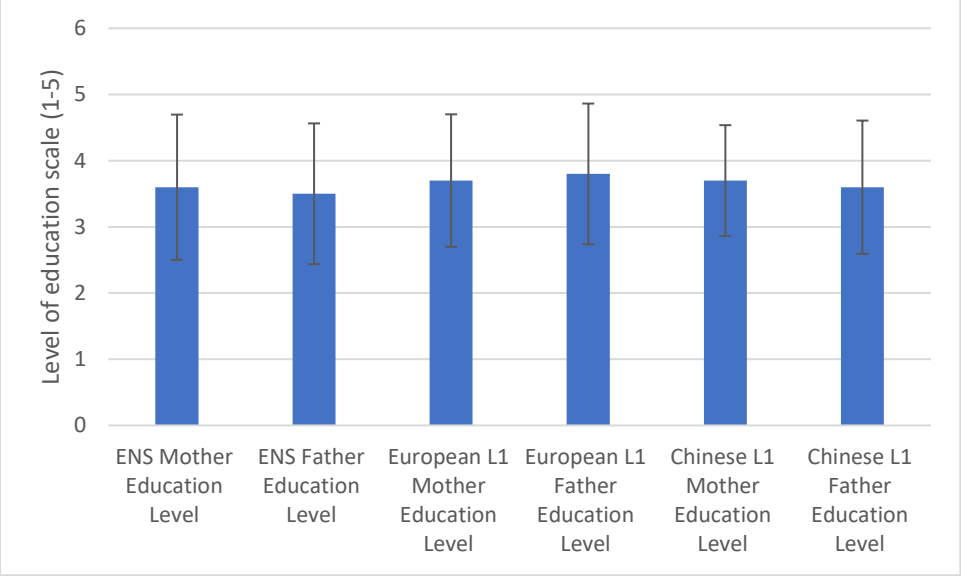

Figure 3. 1 Parental education level

Finally, the most important criterion constituting an independent variable was the first language spoken. Each participant fell into one of the following three language groups: 1) English native speakers, 2) native speakers of one of the European languages, and 3) native speakers of Chinese. Each of those three groups is described in more detail next.

3.2.2 English native speakers

The group of ENS consisted of 59 home students with English as their first language. All the participants in this group claimed English to be their first and dominant language, and almost all of them were born and had gone through the whole education system in the UK. Four of the students in this subsample were born outside of the UK but had been living in

the UK either all their life or since primary school. This group consisted of 43 females and 16 males, with a mean age of 18.49 years (*SD* = .94).

3.2.3 EFL with Chinese L1s

The group of EFL with Chinese L1s consisted of native speakers of Chinese, including Mandarin (*N* = 40) and Cantonese (*N* = 12) dialect speakers. Six participants in this subsample stated their first language to be Chinese without specifying the dialect. The Cantonese speakers were predominantly educated in Hong-Kong which leads to some degree of variability because some primary and secondary school Hong-Kong use English as a language of instruction, next to Cantonese. This means that those brought up in Hong-Kong had more exposure to English in comparison to the participants from the mainland China where Mandarin is the main language of instruction at schools. Nevertheless, both groups: Mandarin and Cantonese speakers can be seen as a homogeneous group in terms of their language and culture, both are very distant from English and European languages included in this study. The whole subsample belonged to the Sinitic branch of the Sino-Tibetan language family. Section 2.6 of the literature review shows how distant these languages are from English and other European languages. This group consisted of 39 females and 19 males, with a mean age of 18.86 years (*SD* = .95).

3.2.4 EFL with European L1s

The group of EFL speaking European L1s comprised 60 students speaking one of the European languages from birth. The final sample included speakers of first languages from several different groupings: Romance languages: Italian, Spanish, French, and Romanian (*N* = 33); Germanic: German, Norwegian, and Danish (*N* = 9); Balto-Slavic: Lithuanian, Polish, Russian, Slovak, and Czech (*N* = 11); and Hellenic (Greek, *N* = 5). One participant in this subsample spoke L1 Finnish and one Hungarian, both languages of the Finno-Ugric family, but they were included for completeness as their presence in the sample did not change the results. Except for the last two, participants in this group spoke languages that are members of the Indo-European language family with a common ancestor.

The make-up of the international student body at the institution where the research was carried out did not allow for a sufficient number of participants to be drawn from the Germanic and Romance families only – i.e., speakers of languages closest to English structurally and/or lexically. Ideally, participants in this group would come from the same country to control for the potential differences in e.g., pedagogical approaches to teaching English, quality and the amount of the English language provision, economic differences that impact the availability of resources such as tutoring by native speakers of English or attending summer schools in the target language country. All these variables increase the variability in this sample and may lead to difficulties in making generalisations. While this group of EFL participants had a higher variability than the EFL participants with Chinese L1s, they nonetheless all came from languages that are structurally and lexically distinctly closer to English than the Sinitic languages. The whole European L1 sample consisted of 41 females and 19 males, with a mean age of 18.67 years (*SD* = .86). See Table 3.1 for a summary of age and gender in all three groups of participants.

Table 3. 1 Summary of participants' age and gender

	Age					Gender			
	N	Mean	Mdn	Mode	SD	Female		Male	
						N	%	N	%
ENS	59	18.49	18	18	.94	43	72.9	16	27.1
EFL European L1s	60	18.67	19	19	.86	41	68.3	19	31.7
EFL Chinese L1s	58	18.86	19	19	.95	39	67.2	19	32.8

3.2.5 EFL students' proficiency level

All the EFL participants were asked to provide their most recent and highest qualification in English and the score obtained. The most frequent qualification was IELTS but some of the students also had TOEFL and Cambridge English. Several students in both groups had obtained an International Baccalaureate (IB) diploma with English as the language of instruction, which is a programme equivalent to A-Levels and prepares students for a university degree. In the sample of 60 EFL students with European L1s, 36 held formal

qualifications in English, and these included IELTS (22), Cambridge English (13), and TOEFL (1). Among the remaining 24 who did not demonstrate any formal proficiency test, 8 held an IB diploma. In the group of 58 EFL students with Chinese L1s, 52 had taken IELTS, while 2 had passed a test other than IELTS (one passed the Cambridge Advanced and one TOEFL). The remaining four Chinese L1 students had not taken any standard language test but had other qualifications (3 had IB and one A-Levels).

The proficiency in English based on standard qualifications in both EFL groups was compared to probe into possible differences at the start of university. To compare the range of language qualifications, the scores from all proficiency tests have been converted into their equivalents in the Common European Framework of Reference (CEFR). See Table 3.2 for conversion conventions and Table 3.3 for the proportion of CEFR bands in both groups. The IB qualifications were also included and set at B2 level in the CEFR (International Baccalaureate, 2022).

Table 3. 2 Proficiency test equivalents

LEVEL	CEFR	IELTS	TOEFL	Cambridge
6	C2	8.5–9.0	–	200–230
5	C1	7.0–8.0	110–120	180–200
4	B2	5.5–6.5	87–109	160–1801
3	B1	4.0–5.0	57–86	40–160
2	A2	N/A	N/A	N/A
1	A1	N/A	N/A	N/A

Note. It was not possible to obtain the highest equivalent for TOEFL.

Table 3. 3 Proportion of CEFR scores in both EFL groups

	A1		**A2**		**B1**		**B2**		**C1**		**C2**	
	N	%	N	%	N	%	N	%	N	%	N	%
EFL European L1s	0	0	0	0	0	0	12	27.3	29	65.9	3	6.8
EFL Chinese L1s	0	0	0	0	1	1.8	37	64.9	19	33.3	0	0

The CEFR bands were further assigned a numeric equivalent with values between 1 and 6 corresponding to CEFR bands of A1, A2, B1, B2, C1, and C2 respectively, with the value of 1 indicating the lowest and 6 the highest level of qualification. The means (calculated on values 1 to 6) for both groups were compared using an independent *t*-test. Homogeneity of variance was assumed for the data, $F(99) = .17$, $p = .684$. According to the *t*-test, a significant difference was found between the EFL with European L1s ($N = 44$, $M = 4.80$, $SD = .55$) and those with Chinese L1s ($N = 57$, $M = 4.32$, $SD = .51$), $t(99) = 4.54$, $p < .001$, $r = .41$, with the group of European students starting university with significantly stronger English than those speaking Chinese L1s (see Figure 3.2).

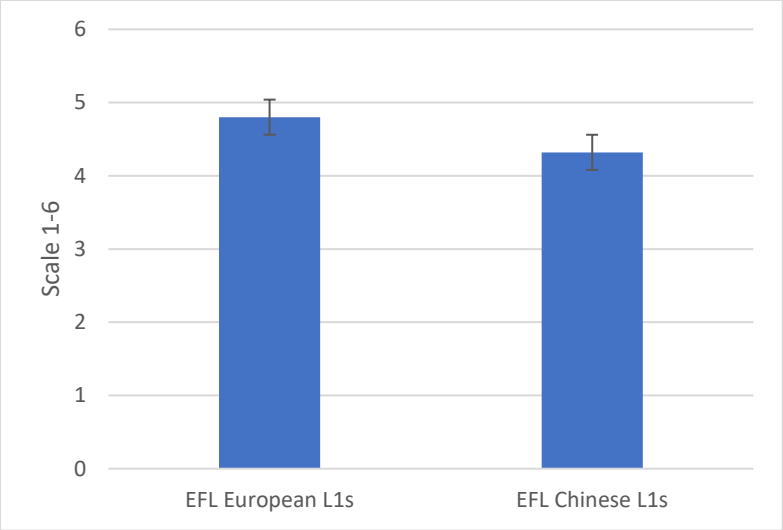

Figure 3. 2 Proficiency level at the start of university in the two EFL groups

The EFL students were also compared on the onset of English acquisition. The group of EFL with European L1s started learning English on average at the age of 7 ($M = 7.15$, $Mdn = 7$, $SD = 2.30$), and the students speaking Chinese L1s started learning English on average one year earlier, at the age of 6 ($M = 6.28$, $Mdn = 6$, $SD = 2.48$). This difference proved to be statistically significant, $t(115) = 1.98$, $p = .025$, $r = .18$, with Chinese L1 speakers starting to learn English significantly earlier in childhood. Despite the fact that participants speaking Chinese L1s started learning English one year earlier than their European counterparts, their level of English as measured by standardised tests prior to entry to university was significantly lower.

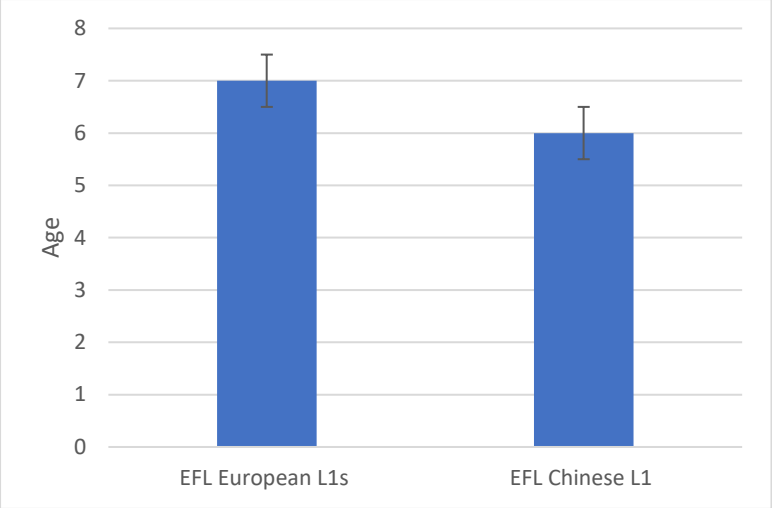

Figure 3. 3 Onset of language acquisition in the two EFL groups

3.3 Time 1 research instruments

A task battery was administered at the start of university to assess cognitive abilities, knowledge of language and literacy, and phonological skills. It was administered again after one year to investigate the magnitude of change. Special attention was paid when selecting the research instruments for the present study, as the same tasks had to be implemented in a heterogeneous group in terms of their first language. Therefore, whenever possible, tasks designed for and validated with both English native speakers and English learners were selected. The tasks are divided into four broad categories, assessing cognitive abilities, knowledge of language, literacy, and phonological skills. In addition, a language background questionnaire was administered, and this is described first.

3.3.1 Background questionnaire

A background questionnaire is a very common tool in linguistic research, as second and foreign language learners comprise a very heterogenous group and 'if such diverse participant background factors are not carefully controlled, this can have varying, and often confounding effects on experimental results' (Sabourin et al., 2016:1). The background questionnaire presented to the participants was designed especially for the purpose of the present study and consisted of two main parts that can be broadly described as demographic and linguistic. The questions were organized in such a way that all participants were able to provide answers to the demographic questions but only the EFL participants

could continue with the second part related to English learning experience, which was not relevant for the ENS. The questions in the first part (for all participants irrespective of their L1) were about participants' age, gender, nationality, length of residence in the UK, caregivers' first languages, and caregivers' level of education. The second part of the questionnaire, administered to the EFL students, asked about the age at which they started learning English, length of stay in any English-speaking countries, their highest qualifications in English with the overall test scores as well as scores for each skill component, whether they were exposed to English at home in their childhood, and whether they had studied using English as the language of instruction. The aim of this questionnaire was to gain a better understanding of the sample and it served for further analyses and exclusions from the study.

3.3.2 Cognitive abilities measures

The cognitive abilities assessed in this study included intelligence and working memory, and this section discusses the instruments used to assess each of these.

3.3.2.1 Intelligence

The concept and importance of intelligence in learning and academic achievement were outlined in section 2.4.5 of the literature review. It was shown that intelligence has a big impact on overall academic attainment in all stages of education, from primary and secondary school (Di Fabio & Palazzeschi, 2009; Laidra et al., 2007) up to the university degree (Busato et al., 2000; Furnham & Chamorro-Premuzic, 2004; Smrtnik Vitulić, & Prosen, 2012). The present study approaches intelligence from the psychometric perspective, treating it as a construct that consists of many abilities that are measurable with a test.

This study assessed participants' fluid intelligence (*gf*). The concept of fluid intelligence originated with Cattell (1943) and refers to 'a purely general ability to discriminate and perceive relations between any fundaments, new or old' (p. 178). Fluid intelligence can be defined as broad ability to reason through drawing inferences and understanding implications. It is therefore best measured in tasks 'that require one to discover the

essential relations of the task for the first time and draw inferences that could not have been worked out before. Tasks intended to measure *gf* should not depend heavily on previously acquired knowledge or earlier-learned problem-solving procedures' (Woodcock & Mather, 1990, as cited in Salthouse et al., 2008:456). The participants should demonstrate 'the ability to solve new problems, specifically the type that are not made easier by extended education or intensive acculturation.… Fluid tasks must involve stimuli and concepts that are about equally available to virtually anyone in a culture' (Kaufman & Kaufman, 1993, as cited in Salthouse et al., 2008:465). It was fluid intelligence that was assessed in the present study because it is independent of knowledge, content, culture, and language, making it a suitable assessment for participants representing a range of academic subjects and, more importantly, a variety of cultural backgrounds and first languages. Fluid intelligence can be assessed with verbal and non-verbal tasks. A non-verbal task was used in this study to avoid confounding intelligence and knowledge of language in the group of EFL participants.

Matrix Reasoning. Fluid intelligence (*gf*) was assessed with the Matrix Reasoning test, a subset of the Wechsler Adult Intelligence Scale (Wechsler, 2011). It is a measure of fluid, non-verbal intelligence, a visual spatial problem-solving task involving a series of figures in which there is a pattern with one figure in the set removed. In this task, the blank must be filled in with an element selected from a list of possible options (see Figure 3.4). The test consists of 26 stimuli that are presented in order of increasing difficulty, with more features distinguishing patterns and more mental processes involved in every consecutive item. The test items were presented in a paper folder with each stimulus printed in colour on an A4 page. The task was not timed but participants were prompted to provide their answers if they spent more than 30 seconds on a stimulus. The test was stopped after three consecutive answers scoring zero. One point was allocated for each correct answer, with a maximum score of 26. The Matrix Reasoning test was selected because its reliability, as reported by the test developer, was very high, with an average among different age groups of alpha = .87 (Abad et al., 2016).

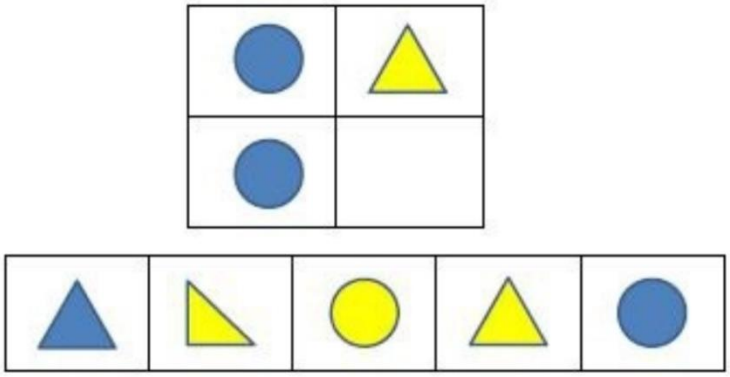

Figure 3. 4 An exemplary item from the Matrix Reasoning test

3.3.2.2 Working memory

Alongside intelligence, working memory is another cognitive ability affecting learning and linguistic processing in the first and second language (Daneman & Green, 1986; Daneman & Carpenter, 1980; Waters & Caplan, 1996). As mentioned in section 2.4.5 of the literature review, the phonological loop is one of the best-studied components of working memory. It is responsible for holding verbal and acoustic information and it is best assessed with verbal digit span tasks. There are two main types of verbal digit span task: forward and backward. In the forward digit span, subjects are required to remember progressively longer sequences of digits. Each string of digits is presented orally, and the participant's task is to repeat them aloud in the same order. This task involves relatively simple processing (such as rehearsal and access of digits) and has only a storage component. A digit span backward, on the other hand, is a task where subjects are presented with digits in strings (similarly to digit forward), but where they are then asked to repeat the digits in the reverse order. It involves a heavier demand on the processing component of working memory, and it is claimed to be a better predictor of linguistic processing than digit span forward (Daneman & Carpenter, 1980). It is discussed in more detail below.

Digit span backward: The task used to assess working memory was digit span backward. This task was designed to assess working memory, one of the executive functions used for temporary storage and manipulation of information. This task combines processing with concurrent storage, as the participant is presented orally with a sequence of digits and their task is to repeat the digits in the reverse order. Therefore, the original order of digits needs to be stored while concurrent processes of organising the digits in the reverse order are

taking place. The task starts with four 2-digit-long string, and the longest string consists of 8 digits. The strings of digits (apart from the first four) come in 2-item sets; that is, there are 2 strings of the same length in each set. One point is allocated for correctly recalling one string and the maximum score is 16 points. The task is discontinued when two strings of the same length are not recalled correctly.

The participants in the present study were administered a verbal digit span backward test in English to test their processing capacities, which means that the EFL were tested in their second language. As explained in the literature review, working memory is not language-independent and participants usually obtain a lower score when tested on this task in their second language rather than in their L1. This is because of the additional cognitive load involved when performing in the second language; it is not that participants have a smaller working memory in their L2s. Ideally, participants should be tested for their working memory capacities in their first language, but this was impossible here as there were many first languages in this study and only a digit span in English was administered. This is, however, justified by the fact that all university students, irrespective of their first language, do pursue their education in English and are assessed at university on the same criteria as English native speakers.

3.3.3 Linguistic knowledge measures

The aspects of linguistic knowledge investigated in the present study are the knowledge of vocabulary and grammar. As stated in sections 2.4.2 and 2.4.3 of the literature review, lexical knowledge is strongly related to reading comprehension (Zhang & Zhang, 2020), writing skills (Stæhr, 2008), and overall academic achievement (Daller & Phelan, 2013). The knowledge of grammar is another marker of general proficiency, and it can predict reading comprehension (Shiotsu & Weir, 2007). When used together, vocabulary and grammar are reliable measures of general proficiency in a given language. This section describes the instruments used for vocabulary and grammar assessment. Given the challenges related to their assessment outlined in the literature review, the tasks used will be critically evaluated to show their suitability in the context of the present study.

3.3.3.1 Vocabulary

Knowledge of vocabulary proved to be one of the best predictors of study success in international students in English-medium universities. It is, however, a construct challenging to investigate due to many ways in which a word can be defined, difficulties in deciding what it means to know a word, and a range of instruments used in this assessment.

Vocabulary Size Test (VST). The instrument used to assess vocabulary was the Vocabulary Size Test. The VST is a measure of written receptive vocabulary size (Nation & Beglar, 2007) available online at https://my.vocabularysize.com/. The construct measured by the VST is written word recognition needed for reading. The VST is a word recognition test where the target lexical item is presented in a sentence providing minimal context and followed by four possible definitions. Participants' task is to choose the best definition of the target word. Figure 3.5 below illustrates an exemplary test item with *miniature* being the target word. It is presented in the following sentence: *It is a miniature*. The only contextual information provided by the sentence is the target word's part of speech. The sentence is followed by 4 possible definitions of that word. The part of speech of the target word is identical to the part of speech represented by the four definitions of the target word. Wherever possible, the words in the definitions were of higher frequency than the item being defined.

It is a **miniature**.

a. very small thing of its kind

b. an instrument for looking at very small objects

c. a very small living creature

d. a small line to join letters in handwriting

Figure 3. 5 Vocabulary Size Test exemplary item

As suggested by its name, the VST is a receptive test of vocabulary size. The concept of vocabulary size in vocabulary knowledge assessment can be criticized on the grounds that it does not reflect lexical knowledge well as the focus is on the number of words rather than how well a word is known. Therefore, the test taker can be rewarded for demonstrating

only a partial or even superficial knowledge of a word. It is argued that a test of vocabulary depth, rather than size, is a better assessment of vocabulary knowledge. Despite these shortcomings, research shows a close relationship between vocabulary size and depth, and both measures seem to be good estimates of the knowledge of vocabulary (Qian, 2002). More importantly, a partial, superficial, or even passive knowledge of a word is still important in this study, as it can be successfully utilised while performing different activities, especially reading. While reading, partial knowledge of a word aids in working out the meaning of that word from the context and facilitates reading comprehension. It is therefore more practical in this context to test the size of vocabulary rather than how well a word is known.

The VST conceptualizes a word in terms of a word family. This means that knowledge of the words *play*, *played*, *playing*, *replay,* and *playful* is treated as knowing one word. The important assumption behind using the word family as a word unit in vocabulary testing is that once the base word is known, *play* in this case, the recognition of other members of the family requires little or no extra effort. The notion of word family in vocabulary knowledge testing can be challenged on the grounds that it assumes that learners have a good knowledge of morphological affixes, and it may not be suitable for students who have not developed this knowledge yet. The test, however, seems to be suitable in the present context as the participants are highly proficient English learners expected to have good knowledge of inflectional morphology.

VST items are based on 1,000 word families from 14 frequency level lists based on the 10 million word token of the spoken section of the British National Corpus (BNC). This means that a total of the 14,000 most frequent word families were the basis for vocabulary selection. The items are ordered according to the level of difficulty, with the most frequent/easiest presented first (Nation, 2006). One of the limitations of the VST is the number of items tested at each frequency level. It is argued that 10 words used to test the knowledge at each frequency level is not enough to accurately estimate vocabulary size in an individual (Gyllstad et al., 2021). Nonetheless, the estimated vocabulary profiles are presented in addition to the raw scores on which the statistical analyses were performed and reported for each language group as a means of comparison against results obtained in

other studies. Nation (2006) suggests the following formula for calculating the estimated vocabulary size:

$$\text{Vocabulary size} = \frac{The\ number\ of\ correct\ answers}{The\ number\ of\ items\ presented} \text{ x frequency levels x 1,000 words}$$

For example, when the correct ratio for the 1st to 14th 1,000-word frequency level test is 50%, the estimated vocabulary size is 50% × 14 × 1,000 = 7,000 words. However, this is not how the software estimates the vocabulary size, because if it were, the maximum would never be higher than 14,000 words, which is not true, as can be seen by looking at the estimated sizes obtained in the test. This suggests that the online tool employs a revised (but not disclosed) algorithm to provide estimates beyond the 14,000 word families. In addition to the raw and estimated vocabulary sizes, the VST software also provides the median reaction time for the correct answer. This value is reported in the final analyses and taps into the speed of processing abilities, a skill that has proved to be very important while performing a variety of activities while at university, as mentioned in some of the studies in the literature review (DeVita, 2002; Zhao & Mawhinney, 2015).

The VST has its limitations, with e.g., a small number of items per frequency level, which may yield inaccurate estimates and large errors in individual learner vocabulary size. Despite these shortcomings, this study investigated group level differences where estimated vocabulary size is a more reliable measure than measures obtained for the purpose of individual learner differences (Gyllstad, 2021). The greatest advantage of the VST for the present study is that the test has been standardised on native speakers of English and English learners with a variety of first languages, such as Vietnamese (Nguyen & Nation, 2011), Japanese (Beglar, 2010), and Russian (Elgort, 2013). It is a reliable and valid test which was examined by Beglar (2010) with a group of Japanese learners of English. It is also very effective in distinguishing between learners at different proficiency levels and has demonstrated generalisability of the test to other population groups.

3.3.3.2 Grammar

Section 2.4.3 of the literature review showed that grammatical knowledge is a construct that is very difficult to define and assess (Shiotsu & Weir, 2007). It also showed the challenges in grammar assessment and tracking the development of grammatical

knowledge in native speakers. In the present study, knowledge of grammar was conceptualised as one of the aspects of general proficiency in English, and students who know more grammatical structures are perceived as more proficient English users. This led to selection of a test based on a wide range of grammatical structures. The choice of a test that would be suitable in the present context was also guided by its suitability for the heterogeneous group of students which included speakers of a variety of first languages and native speakers of English. Taking all these considerations into account, the instrument selected to assess grammatical knowledge was the Which English Grammar Test (Hartshorne et al., 2018).

Which English grammar test. The Which English Grammar Test was originally designed for the purpose of collecting a large pool of data for a study on the critical period hypothesis for language learning (Hartshorne et al., 2018). It is available online at http://archive.gameswithwords.org/WhichEnglish/, but for the purpose of this study it was adapted for administration through Qualtrics. Thanks to that, the actual answers to each question for each participant were recorded and these data allowed calculation of task reliability. The format for selecting the correct answer in the test varies, with 4 main types. The first type of answer format involved picture–sentence matching (Figure 3.6) and there were 6 such questions. Another 2 questions involved a four-alternative forced choice (Figure 3.7). They were followed by 21 fill-in-the-blank questions (Figure 3.8), where participants were asked to select all the answers that could fill the gap out of 4 possible choices. The final 4 questions involved choosing all the grammatical sentences from among 8 possible options (Figure 3.9). Despite this complex answer format, the test is equivalent to a grammaticality judgement task consisting of 132 items of which 95 were critical items. In sum, one point was awarded for each correct judgement of one of the critical items and the maximum score obtainable for this test was 95.

The dog was chased by the cat.

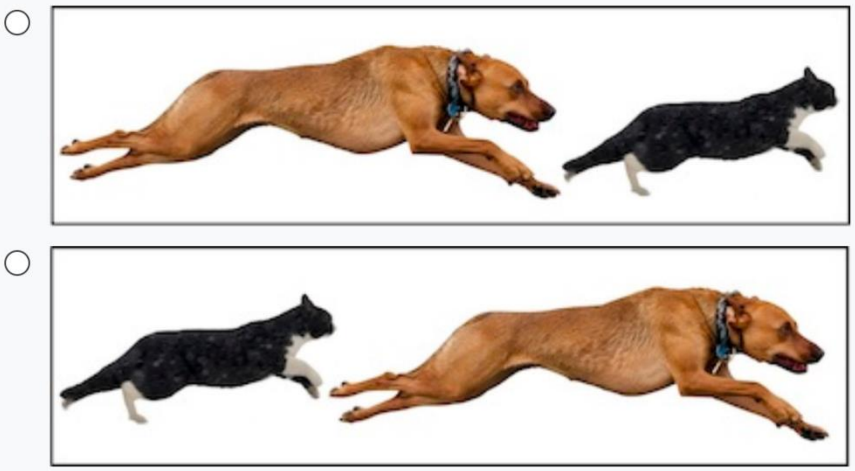

Figure 3. 6 Grammar test picture-matching answer format

Which of the following sentences sounds **most** natural?

- I shan't be coming to the party after all.
- I won't be coming to the party after all.
- Both
- Neither

Figure 3. 7 Grammar test multiple-choice question with one correct answer

I _____ for 6 hours by dinner time.

- will have studied
- will have been studying
- will had studied
- will be studying

Figure 3. 8 Grammar test multiple-choice question with 4 answer options

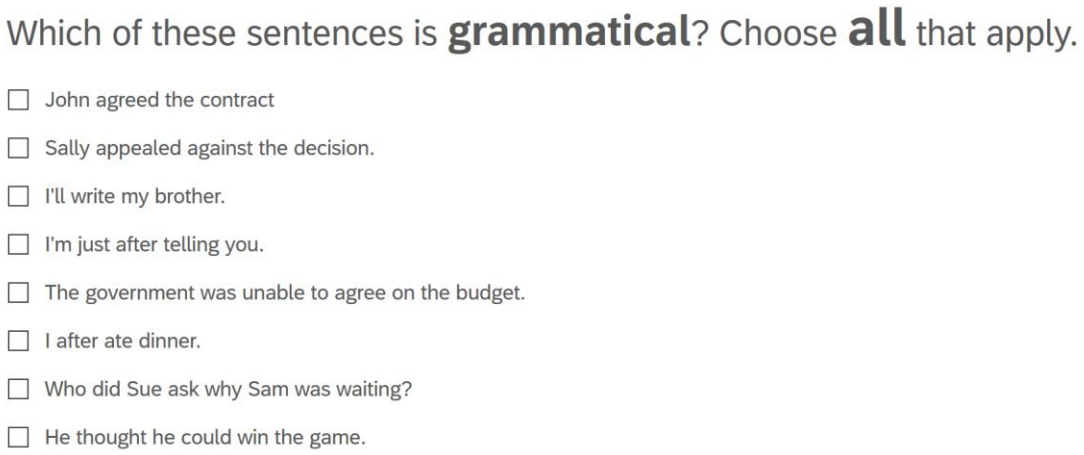

Figure 3. 9 Grammar test multiple-choice question with 8 answer options

The Which English Grammar Test assessed the following grammatical structures: passivization, clefting, agreement, relative clauses, preposition use, verb syntactic subcategorization, pronouns, gender and case, modals, determiners, subject-dropping, aspect, sequence of tenses, and wh-movement. The standardization of the original test was performed on a large sample of 669,498 people at ages between 7 and 89. They represented 38 first languages including English, with around 1,000 native speakers for each L1. It is therefore suitable for native speakers of English as well as English learners representing a range of first languages. The test reliability obtained in the original study proved to be very high (Cronbach's alpha = .86); however, the reliability obtained within the subsample of native speakers of English was much lower (Cronbach's alpha = .66). This low reliability is expected here because, as stated in the literature review, L1 grammar testing is very challenging in native speakers. This test is also convenient to use in this study because it assesses a wide range of syntactic structures within approximately 10 minutes only.

3.3.4 Literacy skills measures

Academic literacy practices – practices of reading and writing within different disciplines – are central in learning and new knowledge development. Writing is also the main mode used in assessing this knowledge, through essays, reports, and written examinations, while at university. They are therefore crucial for ultimate academic achievement. The indices of

reading investigated in this study include reading fluency, comprehension, and speed. Writing skills measures include text length, summarization skills, and spelling.

3.3.4.1 Reading

Good readers are claimed to be skilled in effective word recognition and text comprehension (Hoover & Gough, 1990). Word recognition is conceptualized here as reading fluency which is underpinned not only by accurate but also by rapid word recognition. Reading comprehension is the ability to construct meaning from linguistic input. The speed of word recognition is important because if a reader struggles to recognize words rapidly, pauses frequently, and processes a text too slowly, it taxes cognitive resources, making it difficult to maintain connections within the text and overall text comprehension can be hampered. Once automaticity in word recognition is developed, more cognitive resources are available for text comprehension (Jiang et al., 2012). Research shows a strong relationship between oral reading fluency and reading comprehension (Nathan & Stanovich, 1991; Thurlow & van der Broek, 1997; Wayman et al., 2007). All three aspects of reading: reading fluency, comprehension, and rate, are investigated in the present study.

Sight Word Efficiency. Reading fluency was assessed with the Sight Word Efficiency subtest of the Test of Word Reading Efficiency (TOWRE, Torgesen et al., 2012). This test is an assessment of reading fluency, which is the ability to recognize printed words quickly, and more specifically it measures 'the size of an individual's sight word vocabulary, or words that can be processed as single orthographic units such that they are recognized quickly and with little effort'. The test is used as a 'quick and reliable assessment of word-level reading skills in research studies' (Tarar et al., 2015:320). In this task participants are asked to read 104 real English words as quickly and accurately as they can. The words are organized in increasing level of difficulty and selected based on word frequency, syllabic length and complexity, and number of syllables. The score reflects the number of words read correctly within 45 seconds.

The test is suitable in the present context as it has been designed for the age range of 6 to 24 years 11 months which includes the age range of the participants in the present study.

TOWRE has been standardized on a sample of the US population and proved to be highly reliable as indicated by test-retest reliability coefficient of .92 for the sight word efficiency subset (Torgesen et al., 2012). Studies in different contexts have demonstrated that it is also suitable for other English-speaking populations, and it has been used for research all over the world, including in Australia, Canada, and the UK (Marinus et al., 2013). One of the greatest advantages of this test is that it has been standardized not only on native speakers of English, but also on learners who speak other first languages (Torgesen et al., 2012).

Nelson-Denny reading comprehension test. Reading comprehension, the ability to construct meaning from written linguistic input, was assessed with the Nelson-Denny reading comprehension test (Brown et al., 1993). It consists of 7 reading passages of varying length followed by multiple-choice comprehension-checking questions. The first passage is followed by 8 multiple-choice questions, and each of the subsequent 6 passages is followed by 5 multiple-choice questions. The maximum score that can be obtained is 38 as one point is allocated for every correct answer. The estimated time to complete the task is 20 minutes and the test developer suggested this as a time limit.

Reading comprehension can be assessed in timed and untimed tasks, with the former measuring it under the pressure of time. To understand the construct of reading comprehension in different conditions, some researchers shorten the time limit in the Nelson-Denny test from 20 to 15 minutes (Chateau & Jared, 2000) to see reading comprehension under the pressure of time. This was the case in the present study, so the time limit was set for 15 minutes and the results reflected reading comprehension under time pressure. It was still possible to obtain a measure of untimed reading comprehension: this was achieved through calculating the ratio of correct to total attempted questions, and this second measure reflected untimed reading comprehension. Both measures, for timed and untimed reading comprehension, were included in the final analyses.

The Nelson-Denny test was first standardized on the population of native English speakers in the USA, but its reliability was also explored in a sample of 197 British students (Masterson & Hayes, 2004), and the authors concluded that this test is suitable and a reliable measure to be administered to the student population in the UK. For this reason, the Nelson-Denny reading test has been used extensively in studies conducted outside the USA, including in

the UK (Pasquarella et al., 2012; Shaw & McMillion, 2018) and Canada (Georgiou & Das, 2018). In addition, it has been used to assess the reading comprehension of English learners speaking a variety of first languages, such as Swedish (Shaw & McMillion, 2018) and Farsi (Amirjalili & Jabbari, 2018). It is suitable for the university-age population and assesses comprehension on a variety of passages on different topics, thus reducing the bias that may occur in a sample of participants with a variety of backgrounds and experiences.

3.3.4.2 Writing

Writing skills were conceptualized in the present study as written summarization skills rather than as essay writing skill, used almost exclusively in research elsewhere (Mayor, 2006; Roche & Harrington, 2013). This is because of the nature of the participant sample that consisted of native speakers of English and English learners. English learners were expected to be more familiar with IELTS-style writing tasks and for this reason there was a need to administer a task that would be at a similar level of familiarity to all students. Another advantage of testing summarization skills is that it is genre- and culture-neutral. It does not involve a knowledge of Western rhetoric that could discriminate against those who are less familiar with these writing conventions. A summary is a genre that all the participants were expected to be familiar with to the same extent and which is not over-practised due to test-taking preparation strategies. Students were simply asked to write down as much as they could remember after reading a short text.

Besides summarization skills, other measures used to assess writing skills included the number of words produced and spelling skills. The writing process incorporates the following processes: 1) fluency in generating ideas that can be written down and 2) writing these ideas down before they are forgotten. The second process, also called ‘transcription’, draws on processes involved in retrieving letters and word spelling forms from long-term memory, spelling novel words, and motor planning to produce the letters by hand (Peverly, 2006). More fluent writers are claimed to perform all these tasks at a faster speed so that more cognitive resources can be spent on generating ideas, making their writing of better quality, and therefore, the length of text produced is another indicator distinguishing skilled writers. Spelling is another lower-level skill involved in text writing and consequently better writers are not only faster but also better at spelling (Harrison et al., 2016). All these

measures of writing, namely, summarization, text length, as well as spelling, were obtained by administering one task only, and this was the York Adult Assessment Battery-Revised.

York Adult Assessment-Revised (YAA-R). The main construct measured in the written précis subset of The York Adult Assessment – Revised are summarization skills under the pressure of time. (YAA-R, Warmington et al., 2013). These are 'skills that require complex literacy processing, such as reading a passage, being able to understand it, remember it and then convey the main points in a coherent written form' (p. 54). In this task, participants were first asked to read *The History of Chocolate,* a 492-word nonfiction text (see Appendix G). The summary writing task that followed took 10 minutes; it was a pen-and-paper task in which participants wrote a summary of the source text. The summary was scored based on the content points recalled, and the marking scheme is provided by the YAA-R test developers (see Appendix H). The marking scheme consists of a list of 20 points mentioned in the text. One point was assigned every time a relevant point was mentioned in writing, with a maximum score of 20. Additional measures were obtained by assessing the summary length, and number of spelling errors. Summary length was based on the total number of words produced, and spelling errors were counted in each summary to arrive at the spelling error ratio (a more detailed procedure for handling these measures is provided in section 3.4.5).

YAA-R is a suitable task in this context because it utilizes all the skills that are required in a university learning environment and it measures key literacy skills needed for success in higher education. The test has been standardized on a group of 106 university students in the UK and it proved to have good discriminative power among populations at different levels of writing ability (Warmington et al., 2013). It seems to be suitable for English learners as well because they are assessed on the same criteria as English native speakers.

3.3.5 Phonological processing measures

Phonological processing skills were investigated in this study as the latest research identifies them as important predictors of literacy skills (Schmidtke & Moro, 2020; Trenkic & Warmington, 2019). The two specific phonological skills investigated in this study are

phonological awareness and speed of phonological retrieval. The former can be assessed in tasks involving rhyming, recognizing alliteration, sentence and word segmentation, identifying syllables, and blending letters and words. Task selection for research purposes depends mainly on the participants' age as the task must be aligned with the cognitive abilities of participants. The second skill, speed of phonological retrieval, is measured as the length of time required to name stimuli presented in isolation or serially and is best assessed in rapid naming tasks with four main types of stimuli: letters, digits, colours, and objects. Naming colours and objects is used mainly in research with younger participants. Rapid naming of digits or letters is more suitable for older populations and has more associations with literacy than rapid naming of colours and objects (Bruno & Walker, 1999).

Elision. The first task assessing phonological skills was the Elision task from the Comprehensive Test of Phonological Processing (CTOPP, Wagner et al., 1999). Elision is a test assessing phonological awareness, which is the ability to recognize and manipulate the sounds in a language. Elision measures the ability to remove phonological segments from spoken words to form other words. The test taker is asked to repeat a word with a designated sound deleted. The elision task used in the present study consisted of 20 experimental items divided into two parts. The first part has three lexical items where each needs to be segmented into two separate words or a word and a suffix, for example: 'say *popcorn* without *corn*', or 'say *spider* without -*der*'. The second part consists of 17 lexical items in which a single sound must be deleted, for example: 'say *cat* without /k/', or 'say *winter* without /t/'. All the items were organized in order of increasing difficulty, including the position of the target sound (initial, final, middle), the length of the target lexical item (3–6 words long), the type of sound for deletion, and the complexity of the consonant cluster within which the target sound for deletion is placed (clusters of 1 to 3 consonants). According to the task developer, the task needs to be discontinued after three consecutive scores of zero; however, participants completed the whole set as the original procedure would discriminate international students because the original test was normed on a sample of native speakers of English and follow norms obtained in this group.

The one selected for the present study is from the version designed for individuals in the age range 7 to 24 years old, hence suitable for the participants in this study. The whole task

battery was normed on a sample of 1,656 persons in the USA (Mitchell, 2001). The internal consistency reliability was determined using Cronbach's coefficient alpha. The results showed that it was above .90 for the phonological awareness composite score for almost all age groups. In addition, the confirmatory analysis results suggest that the CTOPP had good construct validity, with factor loadings of .80 for elision (Bruno & Walker, 1999).

Rapid Automatic Naming of digits. The speed of phonological retrieval was assessed with the Rapid Automatic Naming of digits (RAN digits). The version of the task administered in this study is available in the YAA-R task battery. The construct being measured by this task is 'capacity to retrieve phonological codes stored in the long-term memory' (Mitchell, 2001: 58), and these codes may include a variety of concepts, such as digits, as used in the present study. In this task participants are presented with an array of digits organized in 10 rows with 5 digits in a single row (50 digits in total). Their task is to read the digits as quickly and as efficiently as they can. The average number of digits read per one second constitutes the score in this task. Rapid naming of digits was used in this study to avoid confounding performance in this task with English skills in EFL students, as this would be the case if rapid naming of letters was used. Digits are predicted to be learned at the very beginning of second language acquisition and automatized to a greater extent, and therefore the performance in this task will not depend on L2 English knowledge but the pure speed with which participants retrieve the stimuli.

3.4 Procedure

The whole task battery and the background questionnaire were administered across two separate testing sessions. This was motivated by the fact that the time taken to perform all the tasks would take up to 2 hours. Having all the tasks administered within a single session could have imposed fatigue and boredom on the participants, which in turn would impact the reliability of the results (Ackerman & Kanfer, 2009). It was predicted that if organized in two separate sessions, the length of each session would not exceed the time students are expected to spend in for example, a seminar. The vocabulary task was administered as the only task in one of the two sessions because this was the longest task, with an estimated time to complete of up to 1 hour. Thanks to the fact that this task was computerized, it was

possible to administer it to multiple participants at the same time and it was scheduled in group meetings. The remaining 8 instruments and the background questionnaire were administered in another, individual session. The order of tasks within the individual session was set to be the same for all participants, with more cognitively demanding and time-consuming tasks completed first, which is a common research practice (Warmington et al., 2013).

Both sessions were piloted prior to the main study to assess the effectiveness of the task instructions, the reliability of the software in the computerized tasks, participant ID allocation system, and the effectiveness of instructions navigating students into the testing labs, as well as to measure the exact time spent in each session and each task. Two booklets were created and subsequently used in the pilot and later in the main study. The first one was the researcher's handbook with the instructions written down for all the tasks, which ensured that the wording of instructions was identical for all participants (see Appendix F). The second booklet was the answer sheet printed out for each participant where answers for all pen-and-paper tasks were recorded and total scores calculated.

3.4.1 Group session pilot

The group session was piloted in a computer lab on the university campus, the same one where the main data collection took place. This was a sound-proof research facility with capacity for 36 people, equipped with PC stations separated by large screens, making the contact between participants limited. The equipment also included two large projector screens, making it possible to present the task instructions in the form of a PowerPoint presentation. Four students took part in this pilot study: three EFL students speaking Chinese L1 and one native speaker of English. Upon entering the room, each participant was handed a slip of paper with their unique ID and instructions on how to access the computerized Vocabulary Size Test. Participants were let into the room one by one and asked to take a seat at the computer desks. They were asked to read the information page about the study and the data protection information, and to sign the consent form, all of which were waiting for them on the desks. Once all the consent forms were signed and collected, the instructions for the vocabulary task were displayed on the screen. Participants

were informed that they could ask questions before the task began and were asked to leave the room quietly upon answering all the questions in the test. The pilot session demonstrated that the vocabulary test software worked well, participants managed to access the test without any issues, the task instructions were clear, and the system of assigning their IDs worked well. The participants completed the task in varying lengths of time, with a maximum time of 50 minutes.

3.4.2 Individual session pilot

The second, individual session, with all the remaining tasks, was piloted with three participants in a quiet office equipped with a PC. The main data collection took place in another office on the university campus but in similar conditions (a quiet room equipped with a PC). The participants taking part in the pilot study were all EFL students speaking Chinese L1s. As stated in section 2.5.2 of the literature review, Chinese L1 students experience more language-related difficulties and have a slower processing speed than other groups of international students. Therefore, this population served as a baseline in deciding on the clarity and effectiveness of the task instructions. In the individual session, the 8 instruments and the background questionnaire were administered in the following order: 1) Nelson-Denny reading test, 2) Which English Grammar Test, 3) York Adult Assessment-Revised (YAA-R), 4) Matrix Reasoning, 5) Elision, 6) Digit span backward, 7) Rapid Automatic Naming (RAN), and 8) Sight Word Efficiency. The background questionnaire was administered at the very end. The feedback obtained after the individual session led to some improvements in the wording of the instructions for the Nelson-Denny reading comprehension task. One of the statements in the set of instructions (see A below) was improved (see B), as one of the participants in the pilot study was not familiar with the concept of *penalty*:

A) Your score is based on the number of *correct* responses and there is no penalty for incorrect answers.
B) Your score is based on the number of *correct* responses. Since there is no penalty for incorrect answers, it is to your advantage to mark every question you read.

3.4.3 Recruitment

All the participants taking part in the main study were recruited within a single higher education institution. An online screening survey was prepared for all interested in taking part because participation in the study was criteria-based. In particular, recruitment targeted students on linguistically demanding programmes, with the prospective participants being first-year, full-time undergraduate students. They also needed to have as their first language either English, one of the European languages, or Chinese, and to be free from language-related disabilities or problems with sight or hearing. To recruit participants meeting these criteria, two nonprobability sampling strategies were used, namely, convenience sampling and snowball sampling. In convenience sampling, individuals who are relatively easy to identify and contact and who fit the study criteria are recruited (Cohen et al., 2011). The participants were therefore searched for on campus via leaflets, a campaign stall during the Fresher's Fair manned by the researcher, and invitation emails with links to the screening survey sent by the department administrators.

Snowball sampling was used as a parallel method. In this method 'researchers identify a small number of individuals who have the characteristics in which they are interested. These people are then used as informants to identify or put the researcher in touch with others who qualify for inclusion, and these, in turn, identify yet others' (Cohen et al., 2011: 158). The snowball sampling started with participants recruited first through the convenience sampling. They were asked to invite their friends, who in turn spread the word among their friends, so that the recruitment moved forward throughout the whole duration of Time 1 data collection. Email was the main medium of communication with the participants and students were asked to use their official university email in all exchanges. From among the three language groups sought, the one that proved to be the most challenging to recruit from the start was the EFL with Chinese L1s. To maximise the effectiveness of communication with this population, WeChat (a Chinese instant messaging service) was used. It considerably enhanced the speed of communication with students speaking Chinese L1. All the recruited participants signed up for the testing sessions using Doodle, an online scheduling tool.

The recruitment commenced at the end of September 2019. The whole data collection process at Time 1 took place between 30 September 2019 and February 2020, with 90% of the whole sample being tested between October and December 2019.

3.4.4 Testing sessions

Procedures in the main study testing sessions mirrored those outlined for the pilot studies. Each participant attended two sessions: one group and one individual session scheduled within the same week. The number of participants in the group sessions and the length of these sessions varied. Each group session was attended by up to 20 participants and lasted between 30 and 50 minutes. During this session, participants completed only one task, the computerised Vocabulary Size Task. The unique IDs allocated in this session were used in the second, individual session and one year later at Time 2 data collection. During the task, all the participants were looked after and assisted in case of log-in issues. The session was invigilated to ensure that they completed the task independently without conferring with other test takers or using other resources.

In the second, individual session, participants undertook the remaining 8 tasks and the background questionnaire. Upon completion of the second session, each participant was rewarded with £15, asked to sign a slip confirming that the payment was received, and reminded that they would be contacted around the same time in the following year. All the participants attended both meetings in the same order, with the group session preceding the individual one, and all the tasks in the individual session were presented in the same order. They also attended both meetings within the same week. The individual session lasted between 50 and 60 minutes. Both sessions were run by myself.

3.4.5 Task administration

This section describes the procedure of task administration at Time 1 and the procedure for obtaining all the measures for the main analyses. The tasks are presented here in the order of the results presentation in chapters 4 and 5.

Language background questionnaire. The background questionnaire was administered in a conversational form at the end of the individual testing session. The questions were read

out loud to participants and their answers were recorded for them in Qualtrics. This method of administration made it possible to rephrase a questionnaire item for better understanding and students were encouraged to search for their exact language test scores saved on electronic devices carried with them, which strengthened the reliability of the information provided.

Matrix Reasoning. The stimuli were displayed in an A4 paper folder and were printed in colour with one stimulus presented on each page. There were 26 test items preceded by detailed instructions (see Researcher's handbook in Appendix F for the full set of instructions) and two exemplary items, giving the participants an opportunity to practise the task and obtain feedback. The task was not constrained by a time limit, but participants were prompted to move to the next item if they did not provide an answer within 30 seconds of the time of stimuli presentation. The task was discontinued after three consecutive scores of zero. A matrix with all the participants' answers created in Excel, was used to calculate the instrument reliability. The total score obtained by each participant (out of 26) was inserted into a separate Excel spreadsheet for the main analysis.

Digit span backward. In this task, participants were provided with the following instructions:

> I am now going to read lists of numbers. I will stop after each list. When I stop, I want you to say the numbers backwards (in the reverse order). For example, if I say 4–9, you should say: 9–4. I will read each sequence only once, so please, listen carefully. Are you ready?

Two 2-digit-long practice items were presented before the main task began. The main task consisted of 16 strings with 8 sets of the same length, ranging between 2 and 8 digits. The first two sets were 2-digit-long. The stimuli were presented orally at a rate of one digit per second. The task was not time limited, but it was discontinued after two mistakes made within a same-length set. The number of correct answers was recorded manually in the answer sheet. There were two measures obtained for each participant in this task. The first one was the total number of strings correctly recalled. The second measure was calculated

from the length of the last two strings of digits correctly recalled. The numbers of digits in each string were added up and divided by two: for example, if the two last correctly recalled strings were 4 digits long and 5 digits long, these numbers were added up (4 + 5) and divided by two, arriving at a mean score of 4.5. The second measure values were included in the final analysis.

Vocabulary Size Test. The VST was the only task administered in the group session. The following instructions were presented in the PowerPoint before the task began:

> Please switch off your mobile phone. You will do a computer-based English vocabulary test. You will be asked the meaning of 140 English words. Each word will be shown in a sentence, and you will be given 4 possible definitions to choose from. You must not consult with anyone or use a dictionary. The task takes around 30 minutes.
> After you complete the last question (No. 140), something like an error message may appear on screen, with an option *retry* and *cancel*. Press *retry*. You will be asked a couple of language background questions. When you are finished, you'll be given your result. Once you get to the result page, you can close the browser, log off and leave quietly.
> Please go to the website indicated on the slip of paper I've given you. Type the full address, including 'http' (or it won't work). You will be asked for an access code, and participant ID. They are on the same slip of paper. Then follow the instructions on the screen. If you have a question during the test, please raise your hand and I will come to see you.

The results of the VST were recorded by the software and available to download in Excel. The task yielded three measures. The first one was the raw score (the number of correct answers out of 140). The second measure was the estimated vocabulary size based on the raw scores. The software estimates the vocabulary size, expressed as thousands of most frequent word families the test taker is likely to know. The third measure was the median response speed for correct items expressed in milliseconds. The final analyses include the raw scores and the response speed. The matrix of answers for all questions for each participant was used to calculate the test reliability.

Which English grammar test. The grammar test was administered on computer via Qualtrics. The instructions mirrored the original set of instructions provided by the test developer:

> In this quiz, you will decide which sentences are grammatical (correct) and which are not.
>
> Do not worry about whether the sentence is formal or *proper* or is what you learned in school. Scientists have discovered that many of the *rules* taught in school are wrong anyway.
>
> Focus on your gut instincts. Does the sentence sound correct, or does it sound like a mistake -- for instance, a mistake made by a young child or a second language learner?

This task was not timed but it took approximately 10 minutes to complete. Thanks to administration in Qualtrics, the answers to the 95 critical items were downloaded in Excel and scored following the test developer's scoring schedule. The raw score obtained out of 95 points was included in the final analysis.

Sight Word Efficiency. This task consisted of word lists organised in columns and presented to participants in A4 paper format. A list of practice items was presented before the main task began. For the first part, participants were provided with the following instructions:

> I want you to read a list of words as fast as you can without making errors. Let's start with this practice list. Begin at the top, read down the list as fast as you can. If you come to a word you cannot read, just skip it, and go to the next word. Use your finger to help you keep your place if you want to.

The list of 8 stimuli presented in a column on an A4 sheet of paper was given to participants for reading. When they finished, they were provided with instructions for the experimental items:

> OK now you will read some longer lists of words. Read as many words as fast as you can without making errors, until I tell you to stop. OK? You will begin as soon as I turn over the card.

The 104 test words were organised in 4 columns on an A4 sheet of paper. The timer was started as soon as participants started reading the words. Participants were stopped after 45 seconds, and the number of words they read as well as the number of errors was

recorded on the answer sheet. The final analysis was performed on the numbers of words read correctly within 45 seconds. In addition, the time taken, if they finished before the time limit, was also recorded.

Nelson-Denny reading comprehension test. This task was administered on computer and presented in Qualtrics. The participants were provided with the following instructions:

> There are seven reading passages in this reading task. Each is followed by multiple choice comprehension questions. You MUST complete reading each passage first before looking at the questions. Later, you may look back at the material you have read, but do not spend too long on any one question.
>
> When you have completed the questions for one passage, go immediately to the next one.
>
> You have 15 minutes to complete this reading task.
>
> Continue working until you have answered all of the questions or until you are told to stop.
>
> Your score is based on the number of *correct* responses. Since there is no penalty for incorrect answers, it is to your advantage to mark every question you read.
>
> Please do your best to try to answer all the given questions.

Participants were given a 15-minute time limit to complete the task. The results were available to download from Qualtrics in Excel. Two measures were obtained in the task and used in the final analyses. The first one was the raw score for timed reading comprehension out of the maximum score of 38. The second measure was the ratio of correct answers to the total attempted questions, to tease apart speed from reading comprehension. It measured a slightly different construct, untimed reading comprehension.

York Adult Assessment – Revised (YAA-R). In this task, participants read a passage first, and before they started, they were instructed that they would have to write a summary of the main points when they finished. They were provided with the following instructions:

You are now going to read another text but this time you will also write a summary of the main points. I am interested in both: how quickly you can read a text silently, and how much of it you can remember afterwards. This means that when you are writing a summary, you will not be able to look at the text again. OK?

[Present the page with the *History of Chocolate*:] Please read this text silently at your own pace. You should read straight through and not reread sections. While you read, I will be timing you. Please tell me when you have finished so I can stop the timer.

[Remove the page:] OK. You now have 10 minutes to write a summary of the main points. When I say stop, please stop writing; however, if you finish your summary before the time is up, please let me know.

The time taken to read the *History of Chocolate* was recorded and used to calculate the reading speed, one of the measures included in the final analysis. It was expressed in the number of words read per minute and calculated using the following formula:

$$\text{Reading rate} = \frac{Number\ of\ words}{Time\ (in\ s)} \times 60$$

To enable greater efficiency in text analysis, all the handwritten summaries were subsequently typed up in Microsoft Word processing software, retaining all the original wording and preserving spelling errors. In cases where it was not possible to read the handwritten text, the problematic items were removed altogether and not included in analyses. The number of words in each summary was counted, excluding removed items, symbols (+, &, →, *), digits, acronyms (UK, EU), abbreviations (kg, e.g., etc., Mr., approx., PS), and initials. The number of spelling errors was counted for each summary, with each spelling error counted only once; so, if the same word was misspelt in the summary more than once, it constituted only one spelling error. Both American and British English spelling conventions were allowed. The spelling error rate was calculated using the following formula:

$$\text{Spelling error rate} = \frac{Number\ of\ errors}{Total\ number\ of\ words} \times 100$$

Finally, the content of the summaries was analysed for summarization skills, quantified in terms of the number of content points correctly recalled. This procedure was based on the scoring schedule provided by the test developer, consisting of a list of 20 items discussed in the text. One point was awarded every time a relevant point came up in a summary, with a

maximum of 20 points available. To make sure that the marking process was reliable, 20% of the T1 summaries were double-marked by an English language expert. The interclass correlation coefficient (ICC) for the absolute agreement between the two markers was calculated using SPSS. This method evaluates how close the markers were in terms of their scores, i.e., how identical the marks were. In sum, the handwritten summaries provided the following measures of writing used in the final analyses: total number of words produced, error rate per 100 words, and number of content points recalled correctly.

Elision. The Elision task was introduced as a word game to make it more engaging. The following instructions were provided:

> Let's play a word game now. Let me show you how to play it:
>
> Can you say *toothbrush* without *tooth*?
>
> If the participant says *brush*, you say: Good! Let's practise another example.
>
> If the answer is incorrect tell them the right answer and ask to do another example.
>
> Go to the next practice example from the answer sheet.
>
> Ask: Say *airplane* without *plane*.
>
> If the participant says *air*, you say: Good! Let's start the game now.
>
> When arriving at the single sound deletion: The game changes now. I remove one sound rather than the whole word.
>
> Can you say *cup* without *k*?
>
> If the participant says *up*, you say: Good! Let's practice another example.
>
> If the answer is incorrect tell the right answer and ask to do another example.
>
> Go to the next practice example from the answer sheet.
>
> Ask: Can you say *meet* without *t*?
>
> If the participant says *me*, you say: Good! Let's come back to the game now.

There were two practice items before each of the two parts of the task; two practice items before the part involving a word deletion, and two practice items before the second part involving a single sound deletion. The answers to the 20 test stimuli were recorded on the

answer sheet, the total score for each person was calculated manually, and entered in the spreadsheet for analysis.

Rapid Automatic Naming of digits. This task was practised with participants before presenting the experimental stimuli. The practice items consisted of a string of 10 digits running across a page of A4 paper, printed out in a horizontal orientation mode. Participants were asked:

> Please name the following digits aloud from left to right. Please, read as fast as you can but I still need to understand what you are saying. Let's begin with some practice items.

The test items consisted of 10 rows of digits with 5 digits in each row. They were presented on a sheet of A4 paper, printed out in portrait orientation mode. The following instructions were provided before the presentation of stimuli:

> You are going to see some more digits now. Please read from left to right starting at the top row. Please, read as fast as you can but I still need to understand what you are saying. Are you ready?

The timer was set as soon as the participant started reading. The time taken to read the whole list was recorded on the answer sheet and used to calculate the reading rate, expressed in terms of the number of digits read per second. See Table 3.4 for an overview of all tasks and measures.

Table 3. 4 List of all tasks and measures used in the final analyses

Tasks	**Measures**
Matrix Reasoning	Raw score (0–26)
Digit Span Backward	Mean score of two longest strings (0–8)
Vocabulary Size Test	Raw vocabulary score (0–140) Median response time (in milliseconds)
Which English Grammar Test	Raw score (0–95)
Sight word efficiency	Number of words read correctly (0–104)
Nelson-Denny Reading Comprehension	Timed reading comprehension (0–38) Untimed reading comprehension (0–100%)

York Adult Assessment-Revised	Reading rate (no. of words per minute) Total number of words in summary Writing rate (no. of words written per minute) Spelling error rate (no. of errors per 100 words) Summarisation skills (0–20)
Elision	Number of words recalled correctly (0–20)
Rapid Automatic Naming	Number of digits read per second

3.5 Statistical analyses

3.5.1 One-way independent ANOVA

The assessment of data distribution is important because it guides the selection of the correct statistical test. The distribution can be normal, or it can violate normality, resulting in using parametric or non-parametric tests respectively. It is generally advised that the assessment of distribution should not rely on a single method and that several assessments in combination need to be conducted and interpreted in relation to each other. In addition, according to the Central Limit Theorem (CLT), the distribution of sample means approximates to normal distribution, as the sample gets larger. Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold (Field, 2009). Taking all this into consideration, the distribution of data in this study was assessed in two ways: using an eye-ball method and numeric normality tests. Even though normality tests did not account for normal distribution in all measures, the eyeball method showed symmetric bell shapes indicating a normal, Gaussian distribution. What is more, the sample sizes in this study were large enough (ranged between 59-65 participants in each group) for the CLT to hold. This led to selecting parametric tests for all statistical analyses in this study. To increase confidence in the findings (i.e., that by-participant results are not driven by performance on a few atypical items), by-item analysis was performed, which is a supplementary analysis that can be done in ANOVA.

The nature of the first research question involved comparisons between the three groups of participants on all measures administered at T1. To test for the magnitude of between-group differences, a one-way independent ANOVA was performed. This statistical test is used in designs involving one independent variable and at least three groups consisting of different participants. This is what the T1 analyses involved, as there was one independent variable, participant L1, with three levels (ENS, EFL with European l1s, EFL with Chinese), and there were three groups consisting of different participants. The ANOVA test statistic is the *F* ratio, which is the ratio of the amount of systematic variance in the data to the amount of unsystematic variance. It is calculated by dividing the mean sum of squares for between-group effect (MS_M) by the mean sum of squares for within-group effect (MS_R). This can be represented in the following way:

$$F = \frac{MS_M}{MS_R}$$

If the *F* ratio is statistically significant ($p < .05$), it means that at least one difference between means is significant. Therefore, it is necessary to carry out further analyses to find out which groups differ (Field, 2009). These post hoc analyses will be described in more detail below. In addition to the test statistic and significance, another value that helps to interpret each test result is the effect size. Effect size is a quantitative measure of the magnitude of the experimental effect. It indicates whether the experimental effect is small, medium, or large. The effect size can be calculated in a different way for each statistical test and therefore the formula for calculating it will be presented separately for each of the statistical tests described below. The effect size for a one-way ANOVA can be calculated using eta squared (η^2) which is represented by r^2. Eta squared is calculated by dividing the sum of squares in the between-group effect (SS_M) by the sum of squares in the total effect (SS_T). This is represented by the following equation:

$$r^2 = \frac{SS_M}{SS_T}$$

The effect size can be interpreted using the following benchmarks from Cohen (1988):

Table 3. 5 Benchmarks for interpreting effect size

	Small	**Medium**	**Large**
Effect size	0.10	0.30	0.50

3.5.2. Mixed-design ANOVA

In addition to the between-group comparisons conducted at T1, the present study also investigated the extent to which the test scores changed in all three groups across the two time points. In addition to the first independent variable, which was participants' L1, the time at which the participants took the task battery constituted another independent variable, with two levels (T1 and T2). Therefore, this analysis involved between- and within-group comparisons at the same time. In this type of design, a mixed-design ANOVA is used. This test is used in designs involving at least one independent variable with between-group comparisons and at least one independent variable with within-group comparisons. The between-group comparison involved participants' L1 (with three levels), and the within-group comparison involved the time at which the tasks were performed (with two levels). The mixed-design ANOVA yields three sets of results in this study: the main effect of time, the main effect of group, and the interaction between time and group. The main effect of time compares two means: the mean score obtained by all participants at T1 and the mean score obtained by all participants at T2. The main effect of group indicates how the average scores obtained by each group at both time points differ across all three samples. Finally, the time-by-group interaction indicates whether any of the groups improved more than others across the two time points. The effect size for the mixed-design ANOVA (main effects and interactions) was calculated using the following formula:

$$r = \sqrt{\frac{F(1, df_R)}{F(1, df_R) + df_R}}$$

3.5.3 Planned comparisons and post hoc tests

As mentioned above, ANOVAs can only detect the overall experiment significance without the ability to answer the question of which groups differ. It is therefore necessary to run further tests involving pairwise comparisons between every possible combination of groups.

Having three groups of participants would involve a separate comparison between groups 1 and 2 (ENS and EFL with European L1s), groups 1 and 3 (ENS and EFL with Chinese L1s), and groups 2 and 3 (EFL with European L1s and EFL with Chinese L1s). There are two ways to break down the variance accounted for by the model into component parts. This can be done using planned comparison (also known as planned contrast) or a post hoc test. The difference between planned comparison and a post hoc test is linked to the difference between one- and two-tailed tests. Planned comparisons are performed when a specific hypothesis is made before data is collected, whereas post hoc tests are run if no hypothesis exists (Field, 2009). Since both approaches were used in the present study, both are described next.

Planned contrasts enable comparisons between two groups of participants at time. They are used in the present study because, based on the existing evidence presented in the literature review, it is predicted that the group of home students would show stronger language skills than the group of international students in all tasks at the start of university. Therefore, the first comparison was made between the sample of ENS students and the whole sample of EFL students, including both the EFL group with European L1s and the EFL group with Chinese L1s. The second planned contrast involved the two groups of EFL students: the EFL speaking European L1s and the group speaking Chinese L1s. The prediction about the outcome here is based on the existing literature, as Chinese L1 speakers usually demonstrate a lower level of proficiency when compared to other international students (Li et al., 2010; Mayor, 2006). To summarise, there are good reasons to expect that the ENS would obtain significantly better scores in all tasks when compared to the whole group of EFL students, and that participants with European L1s would demonstrate stronger skills in English than students speaking Chinese L1s. Figure 3.10 shows the planned comparisons carried out for the purpose of this study.

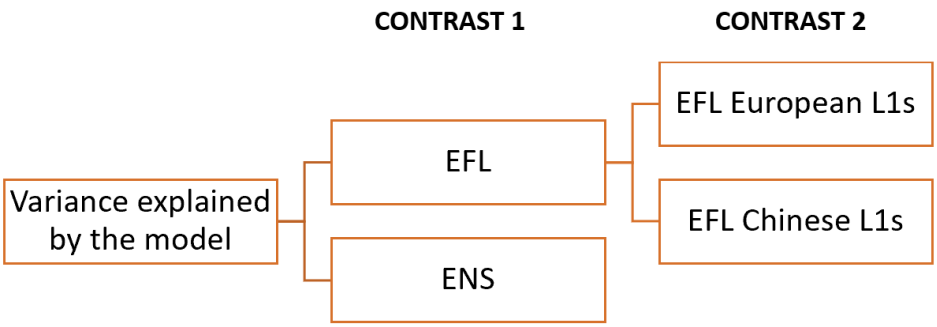

Figure 3. 10 ANOVA planned comparisons at Time 1

As stated above, post hoc tests are post-ANOVA follow-up analyses used where no clear predictions about the direction of difference can be made. This is the case with the comparison between the ENS and EFL speaking European L1s as there is no evidence in the literature to suggest the outcome of this comparison. There are several different post hoc tests and the decision on which one to use depends on the properties of the samples and the distribution of the data. For example, some post hoc tests work better with equal sample sizes and assumed variances (REGWQ and Tukey HSD), whereas others work better in dealing with data sets that may violate the assumption of homogeneity of variance (Games-Howell) and with unequal sample sizes (Hochberg GT2) (Field, Hole, 2002). Since homogeneity of variance was not met in several measures, Games-Howell was used as the main post hoc test and the results yielded there were compared against other tests for consistency.

3.5.4 Independent and dependent *t*-tests

Besides ANOVAs, another statistical test used in this study was the *t*-test, with two variants: the independent and dependent *t*-test. The dependent *t*-test was used in comparisons of two mean scores obtained by the same sample of participants. Dependent t-tests were used as the main analyses investigating language change across the two time points in each of the three groups of students. These analyses involved comparisons of the scores obtained at T1 and T2 by the same group of participants. The independent *t*-test is a parametric statistical test used in between-group comparisons with no more than two groups and different participants in each group. This test was explored in analyses comparing two

different groups of participants. This was the case, for example, when comparing the proficiency level in the two EFL groups of participants, their onset of English acquisition, or their mean length of stay in the UK. The test statistic produced by t-tests is *t*, which constitutes the ratio of the difference between means divided by the estimate of the standard error of the difference between those two-sample means (Field & Hole, 2002). The effect size for the *t*-test is Cohen's *d*, which can be automatically obtained from the statistical software. Cohen's *d* takes the difference between two means and expresses it in standard deviation units. The following benchmarks are used in interpreting Cohen's *d* (Cohen, 1988):

Table 3. 6 Cohen's *d* benchmarks

	Small	**Medium**	**Large**
Cohen's *d*	0.20	0.50	0.80

3.5.5 Task reliability

Task reliability is 'the ability of the measure to produce the same results under the same conditions' (Field & Hole, 2002: 47), and this is a measure of the instrument's internal consistency. This construct in quantitative analysis can have two main forms: the split-half technique and the alpha coefficient. Both calculate a coefficient of reliability that can lie between 0 and 1 (Cohen et al., 2011: 639). Cronbach's alpha was used here as it is one of the most common measures of scale reliability, making the results comparable to other findings. It uses the split-half method which randomly splits the items into two groups in every conceivable way and computes the correlation coefficient for each split. The average of these values constitutes Cronbach's alpha. The value of Cronbach alpha is interpreted using the guidance presented in Table 3.7 below. There are some alternatives to Cronbach's alpha that are claimed to be more reliable and are calculated using more advanced methods (McNeish, 2018), such as ordinal omega (O'Reilly & Marsden, 2021). Nevertheless, Cronbach's alpha is by far the most common instrument reliability index used in L2 research (Plonsky & Derrick, 2016).

Table 3. 7 Reliability of instruments guidance from Cohen et al. (2011: 640)

Threshold	Reliability
> 0.90	very highly reliable
0.80–.90	highly reliable
0.70–0.79	reliable
0.60–0.69	marginally/minimally reliable
< 0.60	unacceptably low reliability

3.6 Assumption checking and data inspection

As stated above, despite the fact that ANOVA is robust even in data sets that are not normally distributed, the data sets were checked for normality. They were also inspected for outliers and missing values. Each of these aspects is briefly described here.

Normality. The notion of normality refers to the way the scores within the experimental sample are distributed. They can be plotted on a histogram with the dependent variable value on the *x*-axis and the frequency of the score on the *y*-axis. The histograms may come in many different shapes, and normal distribution, also called a Gaussian distribution (a bell-shaped curve), is characterized by a symmetrical distribution of scores around the centre. This shape implies that most scores lie around the centre of distribution and the further away we go from the centre, the lower the frequency is. This shape of distribution is characteristic of many naturally occurring phenomena. A distribution can deviate from normal in two main ways, with disruption to symmetry (skew) and pointiness (kurtosis).

The data obtained in this study were tested for normality in two ways: the eyeball method and normality tests. The first method is a visual inspection of a histogram, to see whether it conforms to a Gaussian distribution. Another set of graphs representing the distribution of scores are the P-P (Probability-Probability) Plots that plot the expected values of a normally distributed data in a straight diagonal line and the observed values as individual points. The more closely the points are clustered alongside the diagonal line, the more normally distributed the data tend to be. The plot is created automatically by a statistical package. One of the limitations of visual inspection of a histogram and the P-P plot is the subjectivity

of the judgement. These plots do not indicate whether the distribution is different enough from normal for the assumption of normality to be violated. There are several other ways to capture the distribution numerically and this can be done by running normality tests.

The most popular normality tests are Kolmogorov-Smirnov and Shapiro-Wilk. Both compare the test scores in the sample to a normally distributed set of scores with the same mean and standard deviation. If the result yielded by those tests is not significant ($p > .05$), it means that the distribution of the sample is not significantly different from a normal distribution, and a parametric test can be used for analysis. If, however, the result is significant ($p < .05$), this indicates a significant difference from normal distribution. One of the greatest limitations of normality tests is that it is easy to obtain a significant result in large samples, and it is advised that these tests should be treated cautiously and used in conjunction with the eyeball test.

Homogeneity of variance. Another condition that needs to be satisfied in order to use a parametric test is homogeneity of variance. Variance is defined as the averaged squared deviation between the mean and an observed score. This value is important as it indicates how much a given data point differs from the mean of all data points. It is possible to compare it across all samples which may contain different numbers of observations. The assumption of homogeneity of variance is met when the variance across all experimental conditions does not vary significantly. In cases where the homogeneity of variance is not met, it is possible to use adjustments to correct for a possible error. It is still possible to use a parametric test such as ANOVA; however, a more conservative measurement of the overall significance must be used, such as the Welch and Brown-Forsythe tests.

Outliers. There are several ways in which a dataset can be inspected for the presence of outliers. One of the most common ways is a visual inspection of P-P plots, as nonlinear trends indicate the presence of outliers (Aguinis et al., 2013). Therefore, P-P plots were prepared for all measures and inspected for nonlinear trends. It is advised that visual methods ought to be combined with quantitative methods of outlier identification. This was done by plotting data on boxplots with outliers identified as points that lie beyond the plot's whiskers. There are two types of outliers: soft and extreme cases. Soft outliers are values between 1.5 and 3 interquartile ranges (IQRs). Extreme cases are values that are more than three IQRs above the upper quartile more than three IQRs below the lower quartile.

The approach used for dealing with outliers here is determined by the design of the present study. First, there were three groups of participants compared and therefore outliers were inspected separately for each of the three groups: ENS, EFL with European L1s, and EFL with Chinese L1s. Secondly, the study involved a test battery with 15 different measures of abilities, skills, and knowledge, and it was predicted that some of the students within each group would perform well on some of the tasks and less well on others, depending on individual differences in strengths and weaknesses in different domains of language. The presence of outliers can therefore be interpreted as a natural occurrence due to not being able to perform evenly across several tasks assessing a range of abilities. It was important to detect situation where a single participant produced outliers systematically on most or all measures, which would suggest that a given participant is not representative of the sample and was recruited due to a sampling error (e.g., because of an unreported language disorder in the screening stage), or that the participant was careless while performing the task battery. Taking all these considerations together, the main purpose of outlier analysis was to establish the number of outliers produced by each participant within each group. The outlier analysis was performed separately for each of the two data sets (T1 and T2) and the results of this enquiry will be presented when describing both data sets, the one for T1 in chapter 4 and the one for T2 in chapter 5.

Missing values. Missing values occur when no data value is stored for the variable in an observation. This can be a result of a software malfunction or errors in task administration, and it is important to understand the source of the missing data to handle the analysis correctly. Missing values will be reported for each of the two datasets separately for clarity.

3.7 Chapter summary

In this chapter, the rationale for using a longitudinal study was presented. The chapter discussed participant selection and recruitment, and all the tasks and measures selected to estimate participants' knowledge and skills. In the next chapter, results pertaining to the first research question are presented.

Chapter 4: T1 results

This longitudinal study investigated the knowledge of language and English literacy skills in ENS and international students at a UK university. The aim of this investigation was to better understand the linguistic differences between home students, brought up and educated in English, and international students who were brought up and predominantly educated in a language other than English. The group of international students itself varies linguistically, including speakers of a range of first languages, with some of them being very different from English, such as Chinese, and some more related to English, such as European languages. The purpose was to investigate whether the previous evidence on large and significant differences found for a predominantly Chinese L1 population can be generalised to a different cohort whose first languages are typologically closer to English and who have had a greater amount of exposure to English prior to arriving.

Three groups were recruited at the start of their undergraduate degree to see if language differs between English native speakers, EFL students speaking one of the European L1s, and EFL students speaking Chinese L1s. All three groups were administered a task battery tapping into the knowledge of language (vocabulary and grammar), literacy skills (skills in reading and writing), and other skills that underpin them, such as phonology and language processing skills. All the participants taking part in this study were controlled for their cognitive abilities to rule out the possibility that performance is affected by non-verbal fluid intelligence or working memory. This chapter presents the results obtained in all tasks administered in in-person testing sessions at Time 1 (T1), which pertain to answering the first research question:

RQ1. How much do knowledge of the English language and literacy skills differ at the start of the first year at university between:

a) Home students who speak English as their first language and international students who speak English as a foreign language (EFL)?
b) Those speaking European L1s and Chinese L1s?

Before outlining the results obtained at T1, this chapter first describes the process of data preparation for analysis. This is included in the results section rather than in the

methodology because data preparation was conducted for each wave of data separately and in slightly different ways, and therefore, the description of this activity is presented directly before the results obtained at each time point for clarity. The main part of this chapter consists of the results from all 15 measures obtained in the task battery at T1. This is followed by a brief summary of T1 findings.

4.1 T1 data preparation (outliers, normality, missing values, and task reliability)

The results obtained in the computerised tasks (Vocabulary Size Test, Which English grammar task, and Nelson-Denny reading comprehension) were downloaded automatically to Excel. The remaining pen-and-paper tasks were scored on the answer sheets and the results were entered to Excel manually. The handwritten summaries were typed in Word and analysed using word processor tools such as word count. In the T1 pool of summaries, 5 words from across 3 summaries were removed due to lack of legibility. The inter-rater reliability of the sample marked by another person was very high. The ICC reliability calculated on the overall scores proved to be very high, with α = .97. The data set prepared in this way was inspected for outliers, normality, and missing values, and used for analyses.

Normality: The data were inspected for normality through normality tests first (see Table A.3). According to the Kolmogorov-Smirnov and Shapiro-Wilk normality tests, not all data sets were normally distributed, as indicated by *p* values below .05 in some of the measures. The second analysis involved a visual inspection of histograms for each group and measure. The results of this eyeball technique did suggest that all data sets were normally distributed, as histograms in all groups and measures showed a symmetrical distribution of scores around the centre.

Outliers: The data were inspected for the presence of outliers next. This analysis was conducted by inspecting the P-P (probability) plots for each language group and measure. Upon inspection, none of the P-P plots suggested the presence of outliers, as the observed values indicated a linear trend adhering to the diagonal line of expected values. This method was combined with a more objective method of inspecting outliers on boxplots. The boxplots were inspected for each group and measure, and a total of 47 data points (1.7% of

all data points) were classified as outliers. These included 41 soft outliers (values between 1.5 and 3 interquartile ranges) and 6 extreme cases (values that are more than three IQRs). Following the rationale described in section 3.6, it was found that none of the participants in the ENS group produced more than one outlier. Five participants in the EFL with European L1s produced more than one outlier (4 participants produced two outliers and one produced three outliers). Three participants in the group of EFL with Chinese L1s produced more than one outlier (one participant produced three outliers and two produced two outliers). This shows that none of the participants produced outliers systematically and the existing ones need to be treated as a natural occurrence due to varying levels of knowledge in different domains of language rather than, e.g., a sampling error. Consequently, all the outliers were retained in the final analyses and median scores are reported in addition to each mean score, as the median is more robust to variations in extreme values than the mean (Winter, 2019).

Missing values: Three data points were missing in the whole T1 dataset. These included the vocabulary test score and response time from one participant in the EFL with European L1 sample. This was caused by the vocabulary software's failure to record the test score, probably due to multiple participants saving their answers at the same time within a single testing session. The third data point that was missing was a score in matrix reasoning caused by human error during task administration to a participant from the EFL with European L1 group. It can be concluded that these scores were missing randomly, and their number was not significant in terms of affecting the power of statistical tests. For these reasons, the analyses were performed with these data points missing.

Task reliability. The internal consistency reliability was calculated for instruments with answers represented on a scale. This calculation was performed for the following instruments: Vocabulary Size Test, Nelson-Denny Reading Comprehension, Which English Grammar task, Matrix Reasoning, Elision, and Digit Span Backwards. See Table 4.1 below for an overview of the results obtained.

Table 4. 1 Time 1 tasks' internal consistency

Task	Reliability		
	ENS	**EFL European L1s**	**EFL Chinese L1s**
Vocabulary Size Test	.815	.851	.891
Nelson-Denny reading test	.846	.880	.875
Which English grammar test	.643	.701	.792
Matrix Reasoning	.755	.685	.510
Elision	.754	.765	.663
Digit Span Backwards	.779	.591	.721

The results show that almost all tasks are reliable or highly reliable in all three groups of participants. This does not apply to the results obtained in the grammar test in the ENS group of students, as the result is marginally reliable there. This level of reliability is similar to those obtained in the original study and is expected here as it is difficult to test grammar in native speakers of the language because they will perform close to ceiling. What is more, the population tested in this study is a group of well-educated undergraduate students, making grammar testing even more challenging because this population is exposed to complex grammatical structures in texts read at school and university. A similar problem was identified by other researchers; for example, Devos (2019) used a grammar test in his study as a means of comparing native speakers of English and English learners. The reliability of the grammar task in the group of ENS was so low that this measure was dropped from the study altogether. Apart from the low reliability of the grammar test in the ENS group, all the remaining tasks in the present study obtained reliability measures that are satisfactory even though these tasks have not been standardised on English learners. The vocabulary task was the only one designed for and standardised on native speakers of English and English learners alike and this is reflected in the high reliability scores obtained for this task across the board.

One of the tasks that qualifies for reliability analysis but for which it was not performed was the summarisation skills task. This was a task where participants read a text and then wrote a summary marked against a scoring schedule with one point awarded every time a relevant content point was recalled in a summary and with a maximum of 20 points available. The

difficulty with reliability assessment here stems from the fact that the results may originate in different reading strategies applied by participants: some of them may have concentrated on certain paragraphs, while others may have skimmed the whole text, and these differences in content knowledge may have affected the reliability of the instrument at T1.

Finally, by-item analysis (mentioned in section 3.5.1) was performed on T1 data for the following measures: Matrix Reasoning (scale 0–26), digit span backwards (scale 0–16), VST score (scale 0–140), Which English grammar task (scale 0–95), Nelson-Denny reading comprehension (scale 0–38), written summarisation (scale 0–20), and elision (scale 0–20). The test statistic obtained in these analyses are presented as F_2 to distinguish them from F_1 which is the statistic for the main analysis.

4.2 Results

The data were analysed using SPSS statistical software, version 28.0.0. The analyses of T1 data involved descriptive statistics, one-way independent ANOVAs, planned comparisons, and post hoc procedures (see section 3.5 for descriptions of these methods). The ANOVAs were used to compare the magnitude of the between-group differences, and two planned contrasts were performed to compare the ENS and the whole group of EFL students and also the two groups of EFL students. The initial results obtained in the second planned comparison showed that the group of EFL with European L1s had significantly stronger language skills than the group of EFL with Chinese L1s. Therefore, it was important to compare each EFL group to the group of ENS to understand whether students with European L1s also differ significantly from the ENS students and, if so, on which aspects of the language. To achieve this goal, post hoc procedure was conducted to compare the group of ENS to each of the EFL group. The results obtained in these analyses help to answer the first research question.

The findings for each measure are presented in the following way. The descriptive statistics for each language group are presented first to show the overall picture of the distribution of scores across all three samples. These include the mean, median, standard deviation, and range obtained for each of the three groups. Then, the *F* statistic for the overall significance is reported as F_1 and another *F* statistic (F_2) is reported for the analysis by item. In cases

where homogeneity of variance is violated, an adjusted Welch *F* statistic is reported instead. The results of two planned comparisons come next: the first one for the ENS and the whole EFL sample, the second for comparison between the two EFL groups. To further investigate the between-group differences, pairwise comparisons are reported for the two pairs of participants: ENS vs EFL with European and ENS vs EFL with Chinese L1s. This is to investigate the magnitude of unexplored difference between ENS and EFL with European L1s, as speakers of these languages have not been compared so far. This study also investigates whether the previously observed significant differences between ENS and students speaking Chinese L1s can be reproduced. The results for each measure are represented graphically in the form of box plots.

4.2.1 Intelligence

All three groups performed very similarly on the matrix reasoning task, with the Chinese L1 group performing slightly more strongly than the other two groups (*M* = 21.02, *SD* = 2.24, *Mdn* = 21, Range = 14–25); the ENS students obtained a mean score of 20.66 (*SD* = 3.14, *Mdn* = 21, Range = 13–25), and the EFL with European L1s obtained a mean score of 19.98 (*SD* = 2.89, *Mdn* = 20, Range = 10–25). These differences were not statistically significant. According to a one-way ANOVA, the language group had no effect on performance in the matrix reasoning task, $F_1(2, 173) = 2.080$, $p = .128$, $r = 0.15$, $F_2(2, 75) = .17$, $p = .845$. This means that none of the three groups of participants differ from each other significantly.

Figure 4.1 presents the box plot for the matrix reasoning task and the subsequent box plots follow the same format. To avoid repetition, a detailed explanation for box plots will be provided only once here. Each box represents one group, the horizontal line across each box represents the median, and the 'x' sign within the box represents the mean score. The whiskers above and below each box represent the upper and lower 25% of the data. The solid dots above and below the boxes represent outliers.

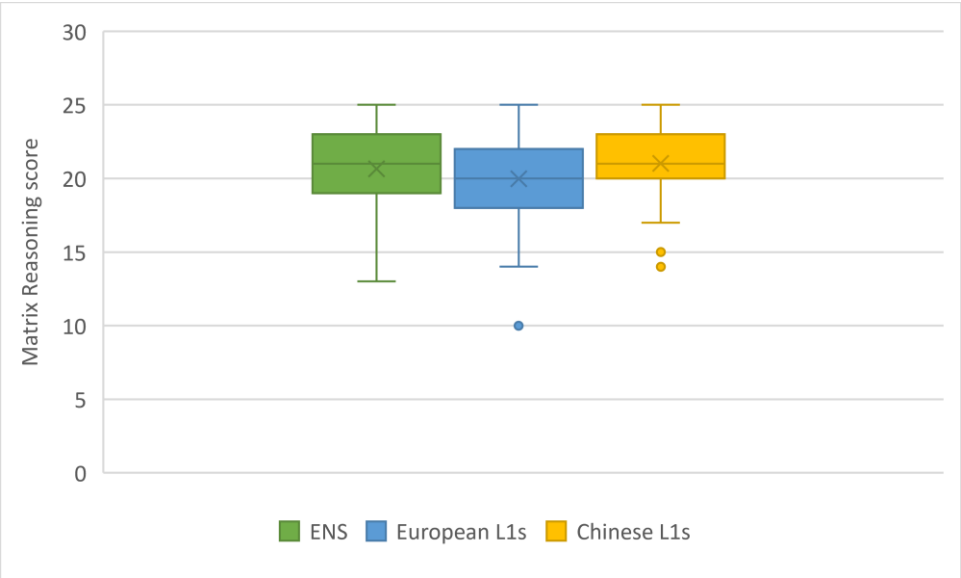

Figure 4. 1 Matrix Reasoning T1 score by language group

4.2.2 Working memory

All three groups performed very similarly on the second cognitive task, digit span backwards, which assessed working memory capacity (see Figure 4.2). The Chinese L1 group obtained a slightly stronger result than the other two groups (*M* = 5.07, *SD* = 1.32, *Mdn* = 4.5, Range = 3–8). The group of ENS obtained a mean score of 4.93 (*SD* = 1.31, *Mdn* = 4.5, Range = 2.5–7.5), and the group of EFL with European L1s obtained a mean score of 4.71 (*SD* = .99, *Mdn* = 4.5, Range = 2.5–7.5). According to a one-way ANOVA, the difference across all three groups was not significant, Welsch $F_1(2, 113.04) = 1.52$, $p = .224$, $r = 0.12$, $F_2(2, 45) = .07$, $p = .936$, meaning that there was no effect of group on performance in the working memory task.

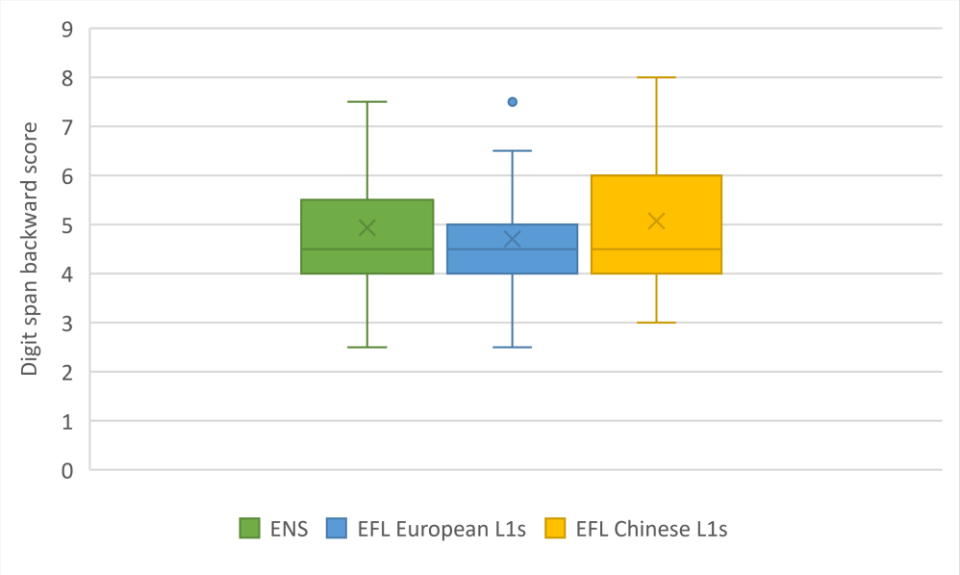

Figure 4. 2 Digit span backward T1 score by language group

4.2.3 Vocabulary knowledge

Statistical analyses were performed on the raw scores obtained in the vocabulary task, where 140 was the maximum score. As shown in Figure 4.3, the group of ENS provided on average 119 correct answers (*M* = 118.75, *SD* = 7.67, *Mdn* = 119, Range = 104–135). The group of EFL with European L1s correctly identified 15 words fewer on average (*M* = 103.85, *SD* = 10.89, *Mdn* = 106, Range = 77–127), and the group of EFL with Chinese L1s knew even fewer tested words than the other EFL group (*M* = 73.76, *SD* = 14.87, *Mdn* = 70.5, Range = 52–114). It is worth noting that the highest score obtained in the group of EFL with Chinese L1 (114) is below the average score in the ENS group (119).

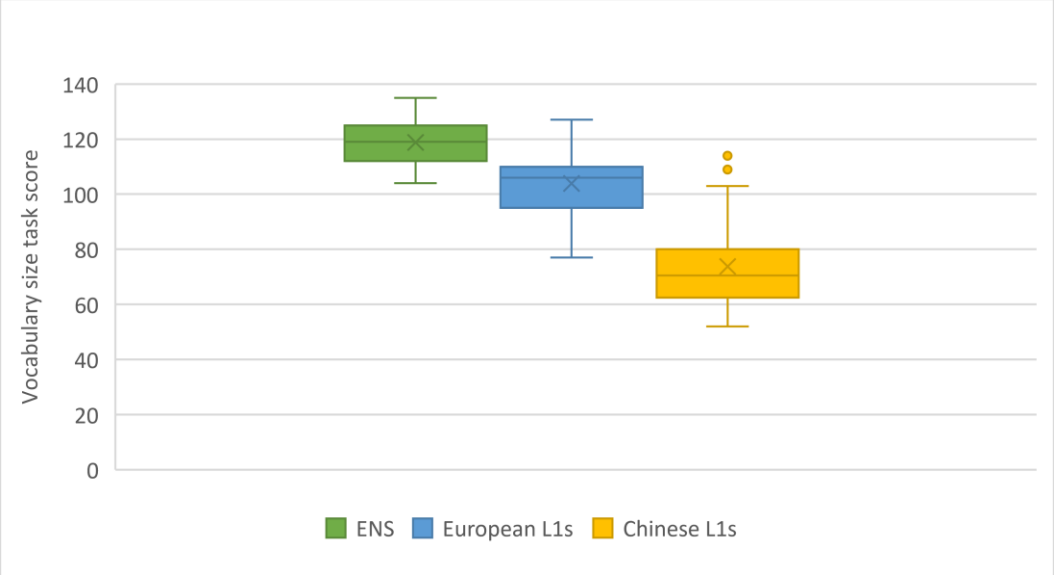

Figure 4. 3 Vocabulary size test T1 score by language group

Levene's test of homogeneity of variance indicated that the assumption of homogeneity of variance has not been met, $F(2, 173) = 8.44$, $p < .001$. According to the Welch's ANOVA, the overall difference was statistically significant, $F_1(2, 107.55) = 214.64$, $p < .001$, $r = .73$, $F_2(2, 276.32) = 61.34$, $p < .001$, $r = .48$. This demonstrates that some groups varied significantly from others. The first planned contrast showed that the group of ENS knew significantly more words than the group of EFL students, $t(160.73) = -19.12$, $p < .001$, $r = .83$. The second contrast revealed that the group of EFL with European L1s obtained significantly higher scores in this task than the other EFL group, $t(104.46) = -12.47$, $p < .001$, $r = .77$. The post hoc analyses further confirmed statistically significant differences, with the group of ENS significantly outperforming EFL groups, both the one speaking European L1s and the one speaking Chinese L1s.

The vocabulary knowledge results described above are based on the raw scores obtained in the task. As stated in the methodology chapter, the VST also provided estimated vocabulary sizes expressed as the number of word families known. The estimated vocabulary scores are reported here (and again in T2 results) to allow comparison of the results obtained in this study with those from other studies. It is also useful to know the estimated vocabulary sizes as they can translate roughly to participants' reading abilities (see section 2.4.2). According to the VocabularySize.com algorithm (based on the raw scores), the average vocabulary size in the ENS group was estimated to be over 16,000 word families ($M = 16{,}415$, $SD = 3{,}896$).

The estimated vocabulary size in the group of students with European L1s was on average over 11,000 word families (*M* = 11,393, *SD* = 2,773), and for the group of students with Chinese L1s it was around 7,500 word families (*M* = 7,441, *SD* = 1,685). These results will be addressed again in the discussion chapter.

4.2.4 Vocabulary test response times

The final measure obtained in the vocabulary test was median response time expressed in milliseconds. The results represent the time taken by participants to read the question and to provide the correct answer. Figure 4.4 shows that the group of ENS provided the correct answer within around 5,000 milliseconds on average (*M* = 5,018, *SD* = 1,052, *Mdn* = 4,899, Range = 2,850–7,951), the group of EFL with European L1s took almost 2,000 milliseconds longer than the ENS students (*M* = 6,790, *SD* = 1,669, *Mdn* = 6,497, Range = 3,768–11,399), and the group of EFL with Chinese L1s took almost twice as long as the group of ENS, as it took them over 9,000 milliseconds to provide the correct answer (*M* = 9,142, *SD* = 2,851, *Mdn* = 8,846, Range = 4,914–16,282). The EFL students with Chinese L1s also demonstrated the greatest variability in their scores. The results for this measure again show that the EFL students with European L1s perform at a much closer level to the ENS students than the EFL students with Chinese L1s do.

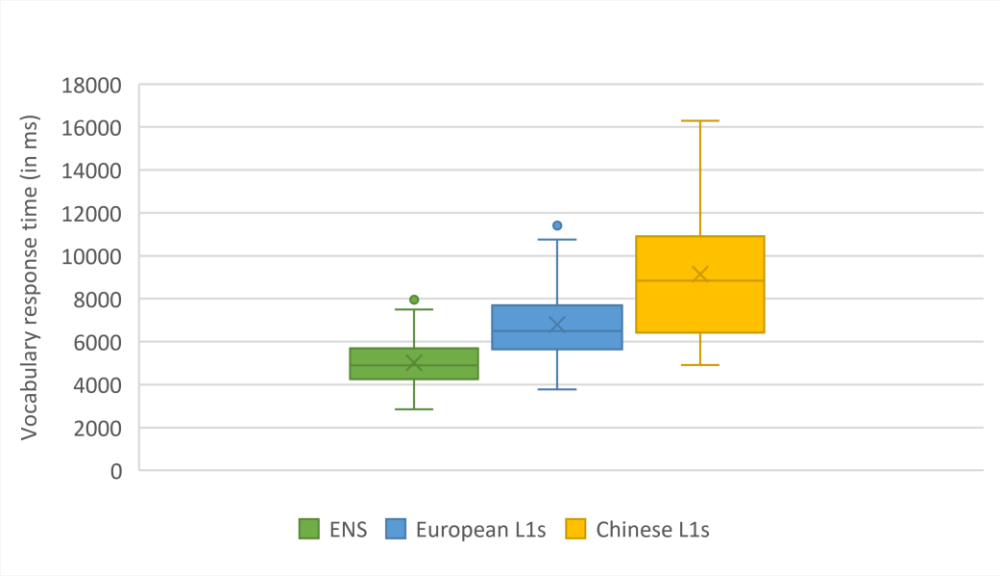

Figure 4. 4 Vocabulary size test T1 response time by language group

The assumption of homogeneity of variance was not met for the data, as indicated by Levene's test, $F(2, 173) = 23.46$, $p < .001$. The between-group differences proved to be significant, according to the Welch's ANOVA, $F_1(2, 102.02) = 66.02$, $p = .000$, $r = .65$, indicating that some of the groups performed significantly differently from others. The first planned contrast showed that the ENS were significantly faster than the group of EFL students, $t(143.41) = 11.51$, $p < .001$, $r = .69$. The second planned comparison further revealed a significant difference between the two EFL groups of students, $t(91.64) = 5.43$, $p < .001$, $r = .49$, with the EFL with European L1s being significantly faster than the other EFL group. According to the Games-Howell post hoc test, the ENS students were significantly faster than both EFL groups, those speaking European L1s and the one with Chinese L1s.

4.2.5 Grammar

According to the results obtained in the Which English grammar test and based on a scale between 0 and 95, the group of ENS performed close to ceiling, obtaining a mean score of 91 ($M = 91.31$, $SD = 2.96$, $Mdn = 92$, Range = 83–95), the group of EFL with European L1s obtained on average 2 points less than the group of ENS ($M = 89.67$, $SD = 3.87$, $Mdn = 90.5$, Range 77–95), and the group of EFL with Chinese L1s' score was on average 11 points lower when compared to the mean for the group of ENS ($M = 80.93$, $SD = 6.73$, $Mdn = 81$, Range 54–92). It is worth pointing out that none of the Chinese L1 participants was able to obtain the maximum score of 95, and the variability in their scores was the highest across all three groups.

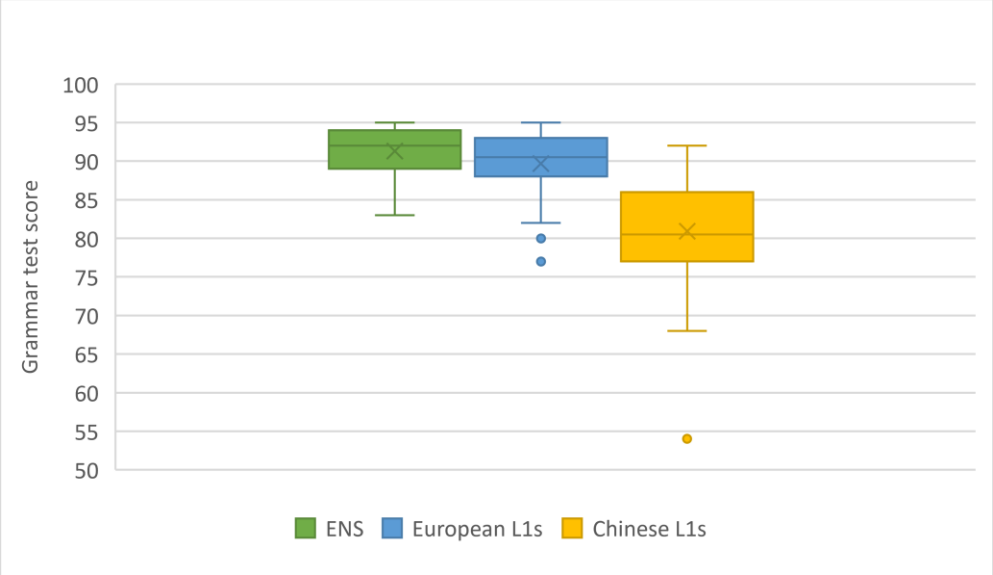

Figure 4. 5 Which English grammar test T1 score by language group

The assumption of homogeneity of variance was not met for the data, $F(2, 174) = 15.32$, $p < .001$, and the overall difference proved to be highly significant, $F_1(2, 107.19) = 57.61$, $p < .001$, $r = .069$, $F_2(2, 174.93) = 19.81$, $p < .001$, $r = .41$. The first planned contrast showed that there was a significant difference between the ENS and EFL students, $t(147.88) = -9.44$, $p = .000$, $r = .62$, with the former outperforming the latter. The second contrast further revealed a significant difference between the two EFL groups, $t(90.30) = -8.61$, $p < .001$, $r = .67$, with those speaking European L1s obtaining a significantly greater score in the grammar test than those speaking Chinese L1s. The results of the pairwise comparisons showed that despite the very small difference in the test scores between the ENS students and the students with European L1s, this difference was statistically significant. The group of ENS also performed significantly better than the EFL with Chinese L1s.

4.2.6 Reading fluency

The scores obtained in the sight word efficiency task indicated the number of words read correctly from a list of 104 vocabulary items within 45 seconds. As can be seen in Figure 4.6, the group of ENS students read an average of 90 words within the allocated time limit ($M = 90.49$, $SD = 9.41$, $Mdn = 91$, Range = 60–104). The group of EFL with European L1s obtained a comparable mean score ($M = 90.82$, $SD = 8.21$, $Mdn = 90$, Range = 66–104). The group of

EFL with Chinese L1s read on average 12 words fewer than the other two groups of students (*M* = 77.95, *SD* = 9.22, *Mdn* = 77.5, Range = 56–98).

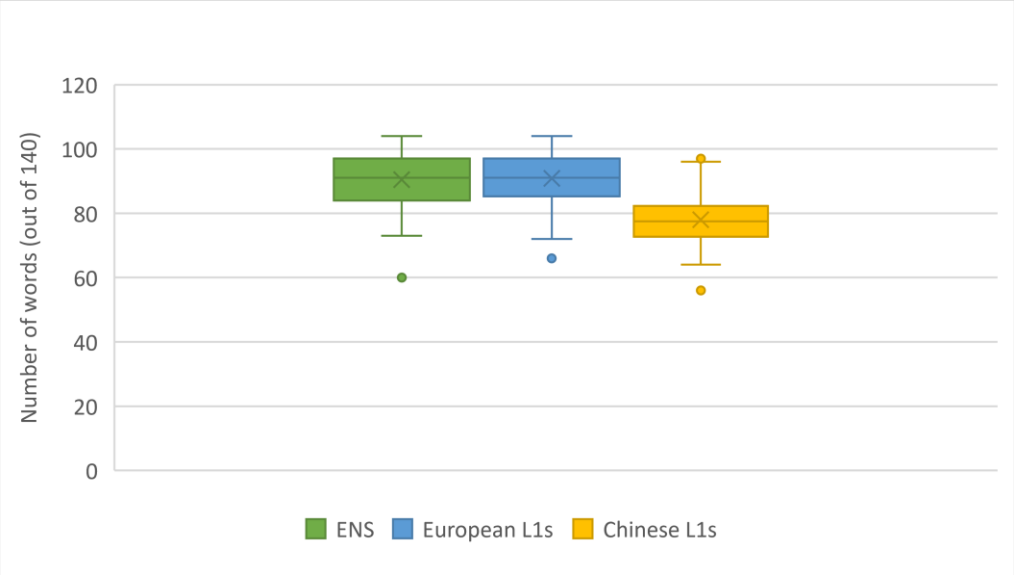

Figure 4. 6 Sight word efficiency T1 score by language group

According to Levene's test, the assumption of homogeneity of variance was met for the data, $F(2, 174) = .37$, $p = .69$. A one-way ANOVA showed that the between-group differences were significant, $F_1(2, 174) = 39.26$, $p < .001$, $r = 0.56$. The first planned contrast revealed that the ENS students read on average more words when compared to the whole EFL group, $t(174) = -4.28$, $p < .001$, $r = .31$. The second contrast further revealed a significant difference between the EFL speaking European and Chinese L1s, $t(174) = -7.80$, $p < .001$, $r = .51$, with the European L1 group reading on average more words than the Chinese L1 group. According to the post hoc procedure, the group of ENS read significantly more words than the group of EFL with Chinese L1s. However, no significant difference was found between the groups of ENS and EFL with European L1s. It is worth to highlight here that, even though the group of EFL with European L1s obtained a slightly higher mean score than the group of ENS, the difference between the group of ENS and the whole EFL sample was still significant. This shows that the results in the EFL sample were driven by the exceptionally low mean scores obtained by the group of EFL with Chinese L1s.

4.2.7 Reading rate

Reading rate was expressed as the number of words read per minute and, alongside reading fluency, is another measure quantifying reading ability. Skilled readers were expected to read faster than less skilled readers, and similarly to the response times for the VST, reading rate is an indicator of the speed of language processing. Figure 4.7 shows that the group of ENS read 222 words per minute on average (*M* = 222.66, *SD* = 77.47, *Mdn* = 217, Range = 100–469). The group of students with European L1s read on average around 50 words fewer than the group of ENS (*M* = 174.38, *SD* = 66.19, *Mdn* = 166, Range = 54–388), and the group of students with Chinese L1s read around 100 words fewer per minute when compared to the ENS participants (*M* = 132.96, *SD* = 51.49, *Mdn* = 130, Range = 46–284). The difference between the group of ENS and EFL with Chinese L1s was twice as great as the difference between the ENS and EFL with European L1s, which means that the EFL with European L1s are right in the middle between the top- and bottom-performing groups when it comes to reading speed.

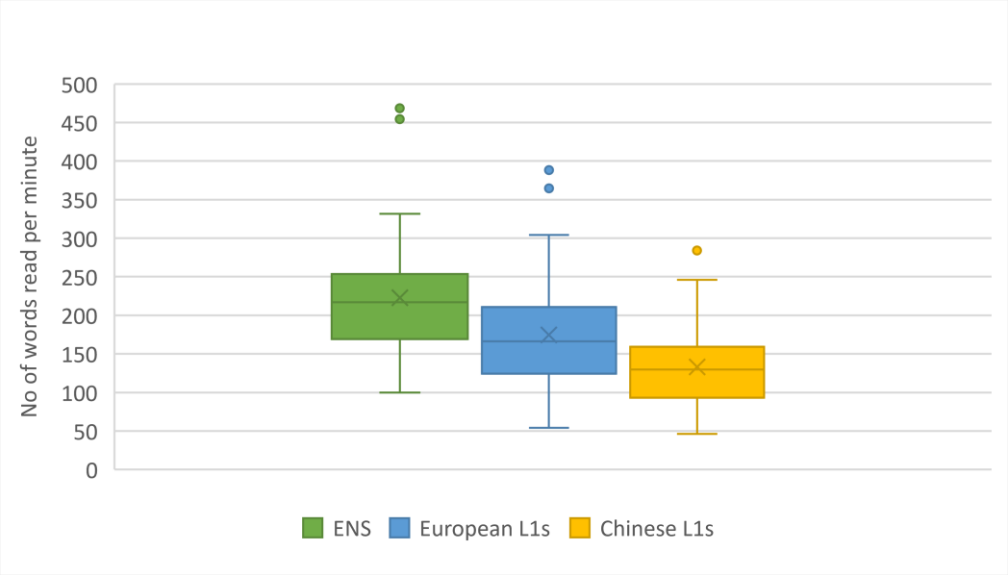

Figure 4. 7 Reading rate at T1 by language group

The assumption of homogeneity of variance was met for the data, $F(2, 174) = 1.89$, $p = .155$. The overall difference proved to be significant, $F_1(2, 174) = 27.09$, $p < .001$, $r = 0.49$. The first planned contrast showed that the group of ENS read significantly more words per minute than the group of EFL students, $t(174) = -6.56$, $p < .001$, $r = .44$. The second contrast further

revealed a significant difference between the two EFL groups of students, $t(174) = -3.41$, $p = .001$, $r = .25$, with those speaking European L1s reading on average significantly more words than those speaking Chinese L1s. According to the post hoc procedure, the ENS students read significantly more words per minute than the group of EFL with European L1s as well as the group of EFL with Chinese L1s.

4.2.8 Timed reading comprehension

Reading comprehension was assessed with the Nelson-Denny test, in which participants read 7 passages and then answered comprehension-checking questions, with the score on a scale between 0 and 38. Good readers were expected to have a better understanding of the text and hence to obtain a higher score for this task. Figure 4.8 shows that the group of ENS participants provided correct answers to 25 questions on average ($M = 25.42$, $SD = 5.99$, $Mdn = 25$, Range 13–35). The EFL participants with European L1s obtained on average 5 points less than the ENS group ($M = 20.37$, $SD = 6.85$, $Mdn = 18$, Range 10–37), and the group of EFL with Chinese L1s obtained a mean score which was about half of that obtained by the ENS students ($M = 12.22$, $SD = 5.82$, $Mdn = 11$, Range = 3–28).

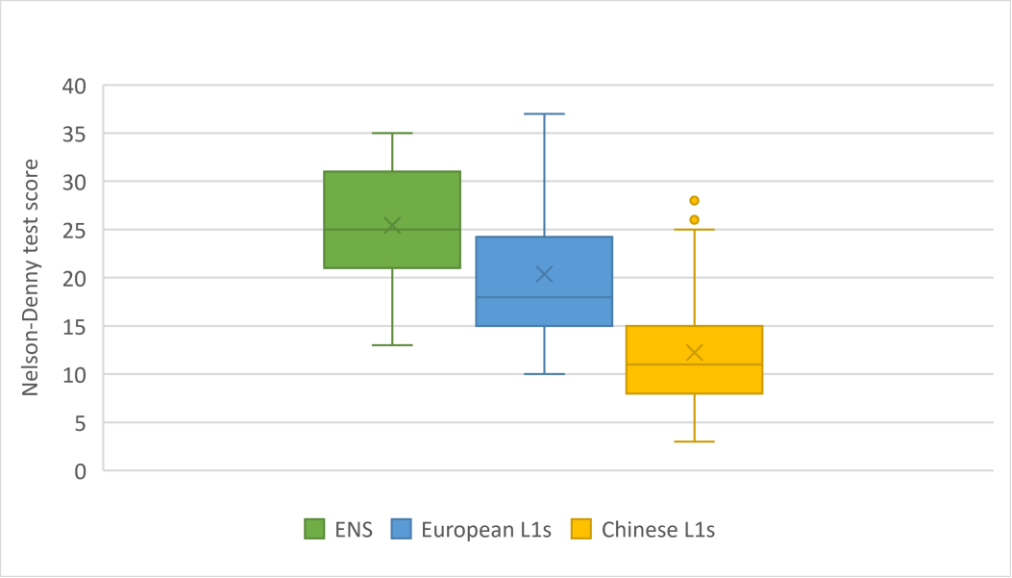

Figure 4. 8 Nelson-Denny timed reading comprehension T1 score by language group

Levene's test indicated that the assumption of homogeneity of variance was met for the data, $F(2, 174) = 1.55$, $p = .215$. A one-way ANOVA showed that the between-group difference was significant, $F_1(2, 174) = 66.57$, $p < .001$, $r = 0.66$, $F_2(2, 111) = 15.55$, $p < .001$, r

= .47. The first planned contrast revealed that the ENS students provided a significantly higher number of correct answers in the comprehension test than the EFL group of students, $t(174) = -9.18$, $p < .001$, $r = .57$. The second contrast further revealed a significant difference between the two EFL groups of students, $t(174) = -7.09$, $p < .001$, $r = .47$, with students speaking European L1s obtaining higher scores than Chinese L1 students. According to the post hoc test, the ENS students significantly outperformed both EFL groups, those speaking European L1s as well as those speaking Chinese L1s.

4.2.9 Untimed reading comprehension

The Nelson-Denny reading comprehension task was timed, with participants being stopped after 15 minutes even if they had not attempted to answer all the questions. Therefore, the scores conflate reading comprehension with reading speed. To exclude the time factor that could have affected reading comprehension, the ratio of correct answers to all the attempted answers in the test was calculated. This value reflects reading comprehension excluding the time factor, which may have had a detrimental effect on the scores obtained. In this second analysis of reading comprehension, the ENS students provided correct answers to over 80% of the attempted questions ($M = .82$, $SD = .10$, $Mdn = .84$, Range = .46–.97). The EFL group with European L1s obtained a slightly lower mean score, with 78% of correct answers out of all the attempted questions ($M = .78$, $SD = .12$, $Mdn = .80$, Range = .55–1.00), and the EFL with Chinese L1s correctly answered 72% of the attempted questions ($M = .72$, $SD = .14$, $Mdn = .72$, Range = .36–1.00). Figure 4.9 visually represents the results for this analysis. In sum, if performed without time pressure, the ENS answered 82% of the attempted questions correctly, and the EFL with European L1s and Chinese L1s 78% and 72% respectively.

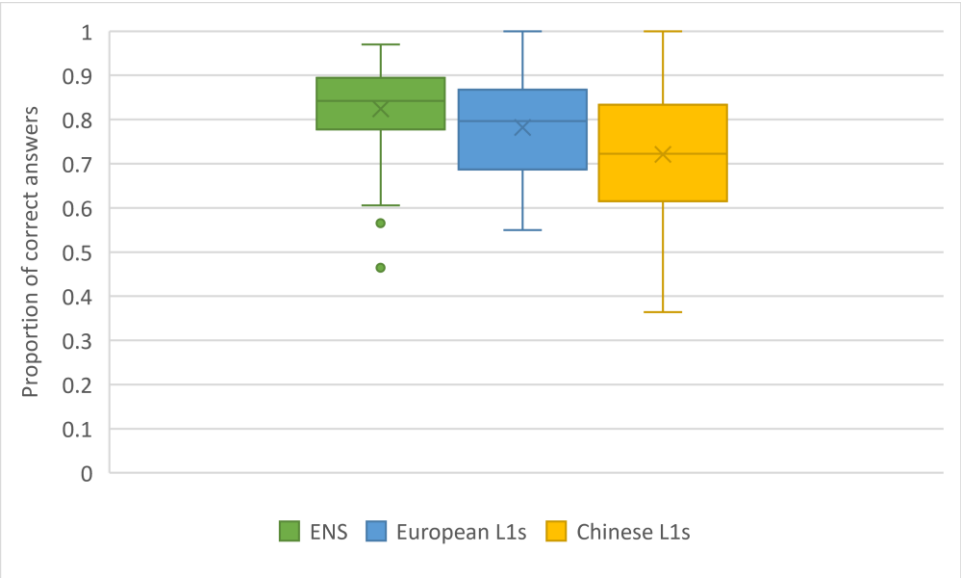

Figure 4. 9 Nelson-Denny untimed reading comprehension T1 score by language group

Further analysis showed that the assumption of homogeneity of variance was met for the data, $F(2, 174) = 3.09$, $p = .05$, and according to the ANOVA, the between-group differences were significant, $F_1(2, 174) = 10.66$, $p < .001$, $r = .33$. According to the first planned contrast, the group of ENS correctly answered a significantly greater proportion of questions than the group of EFL students, $t(174) = -3.74$, $p < .001$, $r = .27$. The second contrast also yielded a significant result for the comparison of the two EFL groups, $t(174) = -2.74$, $p = .007$, $r = .20$, with the group of EFL speaking European L1s answering a significantly greater proportion of questions correctly. According to the post hoc test, the group of ENS did not differ significantly from the group of EFL with European L1s, but they were significantly better than the group of EFL with Chinese L1s. The two analyses of reading comprehension showed that when the speed of reading is taken out of the equation, the difference in reading comprehension between ENS students and EFL students with European L1s disappeared in our sample.

4.2.10 Total number of words

The first measure quantifying writing skills was the total number of words produced in the summaries. It was expected that more skilled writers would produce longer summaries than less skilled writers. Figure 4.10 shows that the group of ENS produced summaries averaging 161 words in length ($M = 161.05$, $SD = 33.19$, $Mdn = 163$, Range = 97–244). The group of EFL

with European L1s produced texts that were on average 12 words shorter in comparison to the group of ENS (*M* = 148.85, *SD* = 39.68, *Mdn* = 153, Range = 83–236). The group of EFL with Chinese L1 produced texts that were on average almost 50 words shorter when compared to those produced by ENS students (*M* = 114.67, *SD* = 33.28, *Mdn* = 111.5, Range = 51–185).

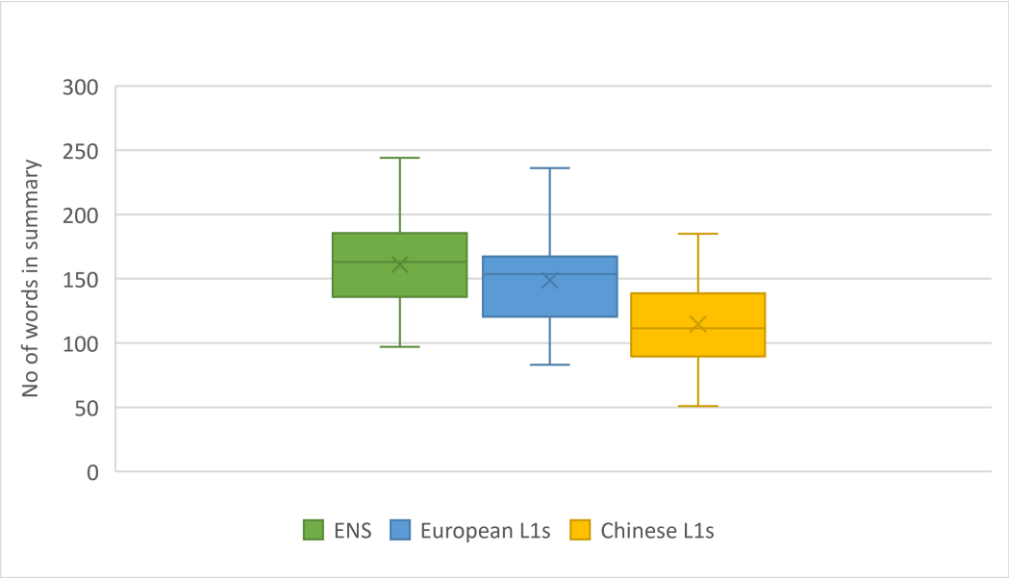

Figure 4. 10 Total number of words produced in T1 writing task by language group

The assumption of homogeneity of variance was met for the data, as indicated by Levene's test, $F(2, 174) = .96$, $p = .384$. A one-way ANOVA showed that the between-group differences were significant, $F_1(2, 174) = 26.71$, $p < .001$, $r = 0.48$. The first planned contrast revealed that the group of ENS wrote significantly more words on average in their summaries than the group of EFL, $t(174) = -5.17$, $p < .001$, $r = .36$. The second contrast further revealed a significant difference between the two EFL groups, $t(174) = -5.22$, $p < .001$, $r = .37$, with the European L1 group producing more words on average than the Chinese L1 group. According to the post hoc test, the ENS students were not significantly different from the EFL with European L1s, but they produced significantly more words in their summaries when compared to the group of EFL with Chinese L1s.

4.2.11 Writing rate

Analysis of the number of words produced in the writing summary has shown that no significant difference existed between the group of ENS students and those speaking European L1s. However, the ENS students did not use the full allocated time of 10 minutes to the same extent as the EFL with European L1s. They completed their summaries on average faster than the EFL with European L1s (530 seconds vs 541 seconds respectively), and a greater proportion of ENS students did not use the full 10 minutes allocated for this task (44% vs 35% respectively). Therefore, writing rate was analysed to find out whether the extra time used by the group of EFL with European L1s proved to be advantageous. Writing rate was obtained by dividing the total number of words in the summary by the time taken to produce it, which reflects the number of words produced per minute. The results, visualised in Figure 4.11, show that the group of ENS produced on average 19 words per minute (*M* = 18.52, *SD* = 3.80, *Mdn* = 18.22, Range = 11–29). The group of EFL with European L1s produced on average 17 words per minute, only two words less when compared to the ENS students (*M* = 16.69, *SD* = 3.81, *Mdn* = 16, Range = 8–24). The group of EFL with Chinese L1s obtained a mean of 13 words per minute (*M* = 13.20, *SD* = 3.82, *Mdn* = 13.30, Range 6–24).

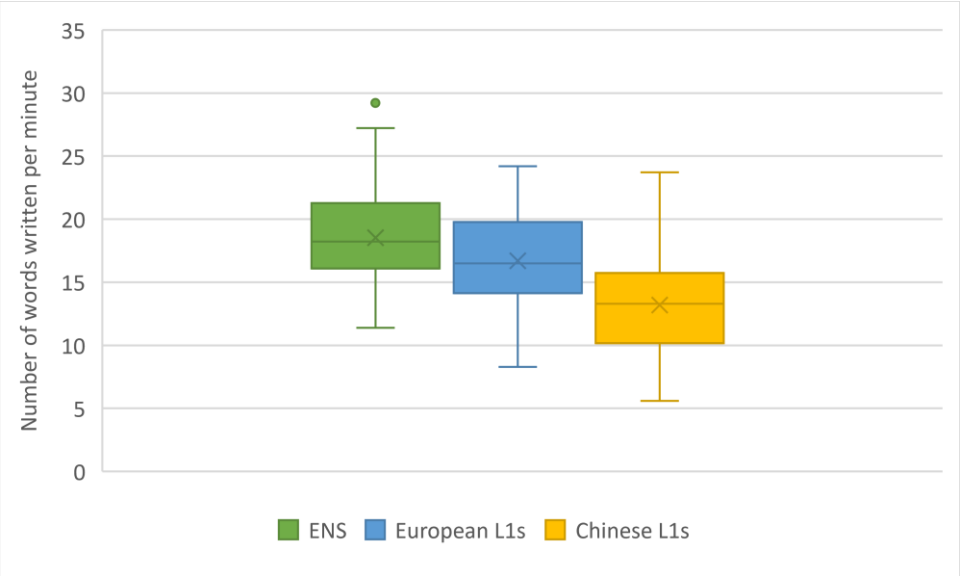

Figure 4. 11 Writing rate at T1 by language group

In this task the assumption of homogeneity of variance was met for the data, $F(2, 174) = .02$, $p = .984$, and the overall difference proved to be significant, $F_1(2, 174) = 29.49$, $p < .001$, $r = .50$. According to the first planned contrast, $t(174) = -5.89$, $p < .001$, $r = .41$, the ENS students produced significantly more words per minute than the EFL group. The second planned contrast showed that the difference between the two EFL groups was also significant, $t(174) = -4.99$, $p < .001$, $r = .36$, with European L1 students demonstrating a greater writing rate than the Chinese L1 participants. The post hoc test showed that the ENS students wrote significantly more words per minute than each of the EFL groups. This extra analysis of writing rate shows that the extra time used by the EFL with European L1s proved to be facilitating in terms of writing. These findings have an important pedagogical implication and this will be addressed again in the discussion chapter.

4.2.12 Summarisation skills

The written work was analysed not only in terms of quantity, but also in terms of quality, by assigning points every time a relevant information point from the story read beforehand was mentioned in the summary. Twenty such content points were selected by the test developers and included in the scoring schedule; participants could obtain a maximum of 20 points. According to Figure 4.12, the group of ENS recalled on average 8 content points in their summaries ($M = 8.37$, $SD = 2.65$, $Mdn = 8$, Range = 3–15). The group of EFL with European L1s recalled on average one point fewer than the ENS students ($M = 6.97$, $SD = 2.54$, $Mdn = 7$, Range = 2–13). The group of EFL with Chinese L1 produced half as many content points when compared to the ENS students ($M = 4.31$, $SD = 2.13$, $Mdn = 4$, Range = 0–12).

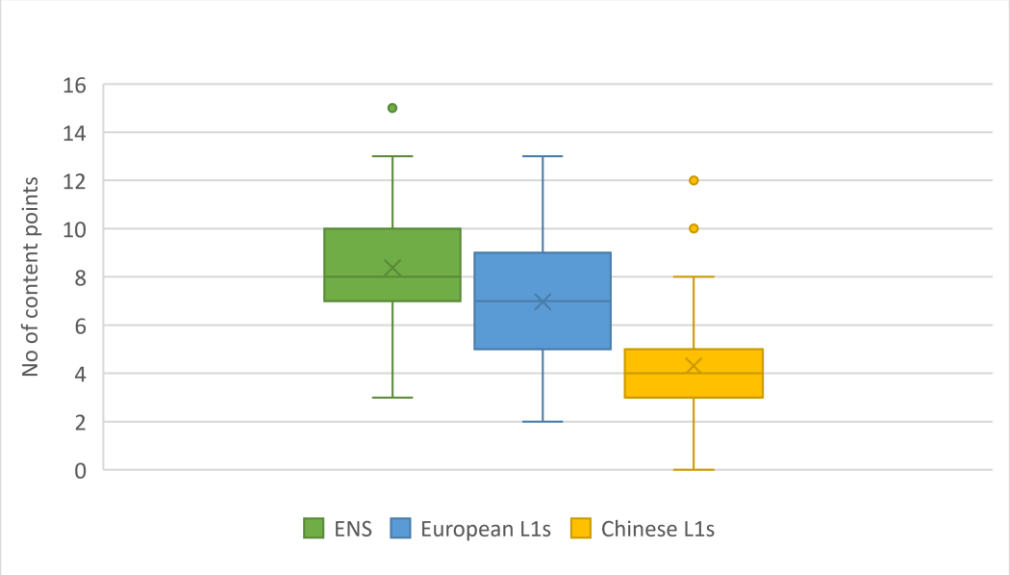

Figure 4. 12 Summarization skills at T1 by language group

The assumption of homogeneity of variance was met for the data, as indicated by Levene's test, $F(2, 174) = 1.89$, $p = .154$. A one-way ANOVA showed that the between-group differences were significant, $F_1(2, 174) = 41.44$, $p < .001$, $r = 0.57$, $F_2(2, 57) = 3.92$, $p = .025$, $r = .35$. The first planned contrast revealed that the group of ENS recalled significantly more content points in their summaries when compared to the whole EFL group, $t(174) = -7$, $p < .001$, $r = .47$. The second contrast further revealed a significant difference between the two EFL groups, $t(174) = -5.89$, $p < .001$, $r = .40$, with the group speaking European L1s obtaining significantly higher mean scores when compared to the other EFL group. According to the post hoc test, the ENS students obtained a significantly higher mean in the summarisation task than each of the EFL groups of students, those speaking European L1s and those with Chinese L1s.

4.2.13 Spelling

Finally, the handwritten summaries were analysed in terms of accuracy in spelling, with the number of spelling errors per one hundred words constituting the spelling error rate. As shown in Figure 4.13, the group of ENS obtained a mean spelling error rate of .94 ($M = .94$, $SD = .79$, $Mdn = .88$, Range = 0–2.86). The EFL students with European L1s obtained a higher spelling error ratio when compared to the ENS students ($M = 1.36$, $SD = 1.31$, $Mdn = 1.15$,

Range 0–7.10). The group of EFL with Chinese L1s made more than twice as many spelling errors as the ENS students (*M* = 2.40, *SD* = 2.00, *Mdn* = 1.84, Range = 0–9.09).

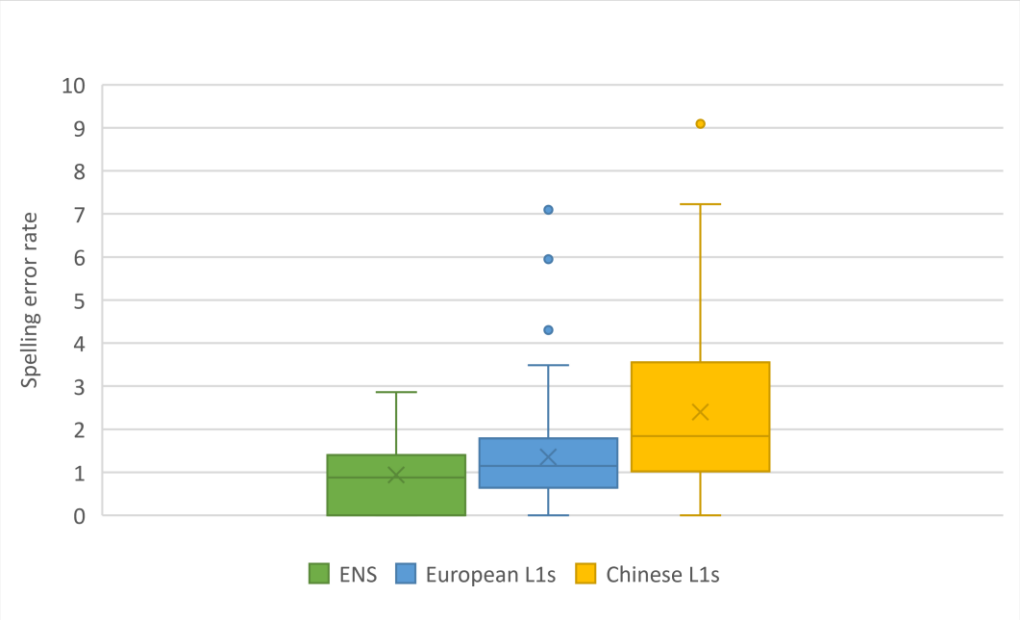

Figure 4. 13 Spelling error rate at T1 by language group

The assumption of homogeneity of variance was not met for these data, $F(2, 174) = 13.66$, $p < .001$, but the Welch's ANOVA proved to be significant, $F_1(2, 102.88) = 13.89$, $p < .001$, $r = 0.39$. The first planned contrast showed that the group of ENS made significantly fewer spelling errors when compared to the EFL sample, $t(152.43) = 5.01$, $p < .001$, $r = .37$. The second contrast revealed that the EFL with European L1s made significantly fewer spelling errors than the group with Chinese L1s, $t(97.58) = 3.33$, $p = .001$, $r = .32$. According to the Games-Howell procedure, the ENS students were not significantly different from the EFL with European L1s, but they were significantly better in spelling when compared to the group of EFL with Chinese L1.

4.2.14 Elision

Elision is the first of the two tasks assessing phonological skills in English, where participants were asked to repeat a word after a sound deletion. The test scores were based on a scale between 0 and 20. As seen in Figure 4.14, the mean score in the groups of ENS and EFL with European L1s was almost identical: both groups provided almost 18 correct answers on average. The mean was 17.56 in the group of ENS (*M* = 17.56, *SD* = 2.20, *Mdn* = 18, Range =

12–20) and 17.63 in the group of EFL with European L1s (*M* = 17.63, *SD* = 2.34, *Mdn* = 18, Range = 9–20). The group of EFL participants with Chinese L1s provided on average 2 correct answers fewer than the other two groups of students (*M* = 15.50, *SD* = 2.66, Mdn = 16, Range = 8–20).

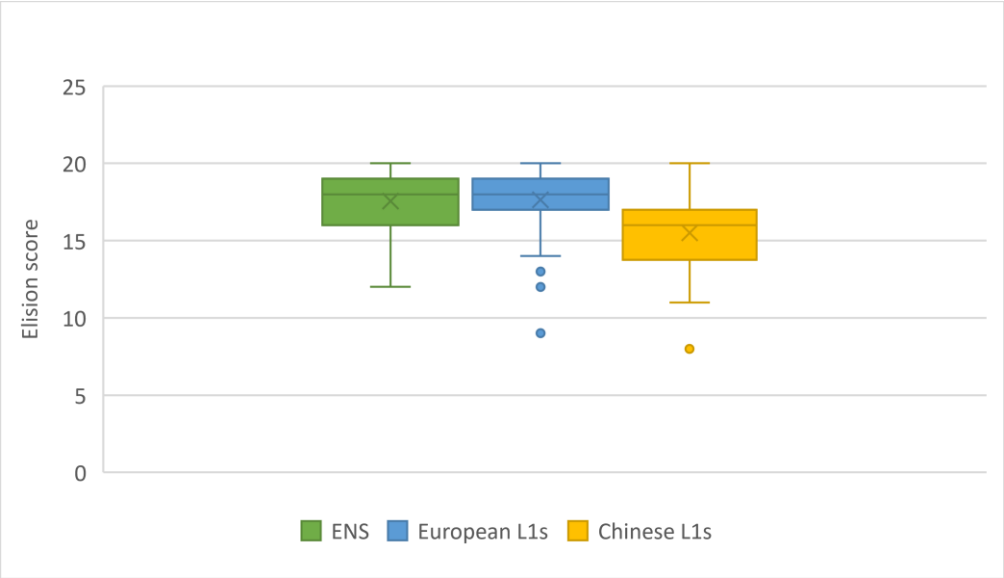

Figure 4. 14 Elision test score at T1 by language group

The assumption of homogeneity of variance was met for the data, $F(2, 174) = 1.26$, $p = .287$. A one-way ANOVA showed that the between-group differences were significant, $F_1(2, 174) = 14.85$, $p < .001$, $r = 0.39$, $F_2(2, 57) = 1.97$, $p = .148$. The first planned contrast showed that the group of ENS correctly repeated significantly more words than the group of EFL, $t(174) = -2.59$, $p = .010$, $r = .19$. The second contrast revealed a significant difference between the two EFL groups, $t(174) = -4.82$, $p < .001$, $r = .34$, with European L1 speakers demonstrating stronger phonological awareness than the group of Chinese L1 speakers. According to the post hoc procedure, the ENS participants were not significantly different from the EFL with European L1s, but their results were significantly better when compared to the other EFL group.

4.2.15 Rapid automatic naming of digits

The rapid naming of digits is the second phonological measure and the last measure in the whole task battery presented in this chapter. In this task, participants were asked to read a

list of digits arranged in rows. The result is the reading rate, i.e., the number of digits read per second. As seen in Figure 4.15, the ENS participants read on average almost 3 digits per second (*M* = 2.79, *SD* = .50, *Mdn* = 2.78, Range = 1.72–4.17). The EFL participants with European L1s read around 2.5 digits per second, slightly less than the ENS students (*M* = 2.58, *SD* = .47, *Mdn* = 2.50, Range 1.61–3.85). Finally, the group of EFL with Chinese L1s read almost 2.5 digits per second, slightly less than the EFL with European L1s (*M* = 2.49, *SD* = .70, *Mdn* = 2.33, Range = 1.56–5.56).

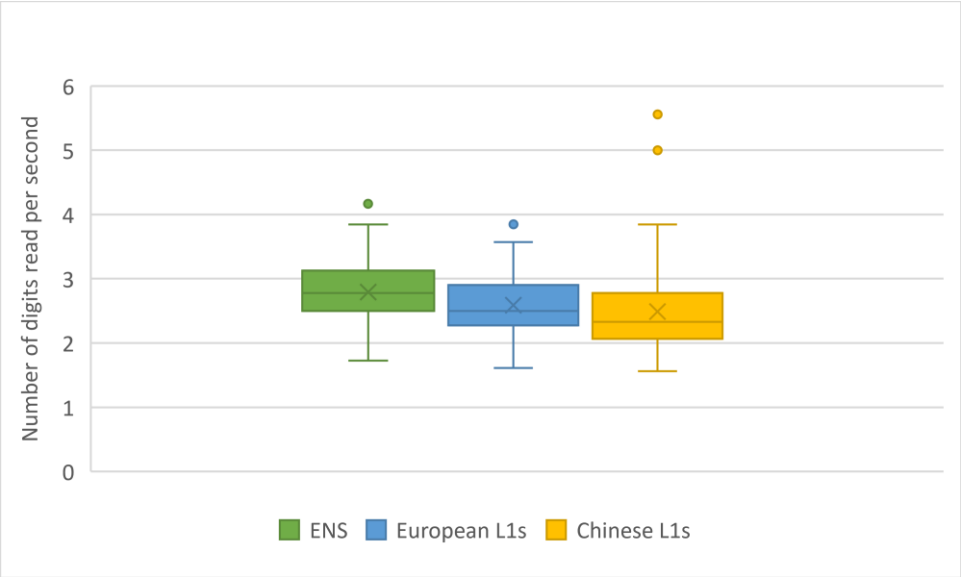

Figure 4. 15 RAN digits T1 score by language group

The homogeneity of variance was met for the data, $F(2, 174) = 2.01$, $p = .137$, and the overall difference proved to be significant, $F_1(2, 174) = 4.15$, $p = .017$, $r = .22$. The first planned comparison yielded a significant result, $t(174) = -2.74$, $p = .007$, $r = .20$, demonstrating that the group of ENS could read significantly more digits per second than the whole EFL group of students. The second planned comparison showed that there was no significant difference between the two EFL groups in the retrieval task, $t(174) = -.93$, $p = .355$, $r = .07$. It is worth stressing that this is the only measure where no significant difference was found between the two EFL groups of students. According to the post hoc procedure, the group of ENS and EFL with European L1s were indistinguishable but the group of ENS performed significantly better than the group speaking Chinese L1s.

4.3 Summary of Time 1 findings

The results obtained in the cognitive tasks showed that there were no differences between the three groups in this study in terms of non-verbal fluid intelligence and working memory. Despite that, the results show large and significant differences in *all* measures of knowledge of language, literacy, and phonological skills between the group of native speakers of English and the group of English foreign language learners. This is confirmed by the results of the first planned contrast for all measures (see Table A.1 in Appendix A). This means that these two groups differ in all indices of language knowledge, including knowledge of vocabulary and grammar. There are significant differences in literacy skills such as reading and writing, and in measures that underpin these skills: phonological skills, speed of processing, spelling, and reading fluency. These results are consistent with the large body of evidence showing that ENS students have stronger language skills than international students at the start of university courses.

The results obtained in this study also corroborate the existing evidence on the magnitude of differences between native speakers of English and those speaking Chinese L1s. These two groups proved to be very different in all measures. The sample of British home students significantly outperformed the group of students speaking Chinese L1s in all tasks, demonstrating significantly better knowledge of vocabulary and grammar and significantly better skills in reading, writing, and phonology. It is worth stressing that the sample speaking Chinese L1s obtained slightly higher mean scores in the cognitive tasks when compared to the group of ENS, but these abilities do not seem to be facilitating when performing in the linguistic part of the task battery.

This study presents findings on previously unexplored differences between two groups of EFL students that are controlled for the first language spoken. The results show that there were significant differences between the two EFL groups, with the European L1 students performing more strongly on all linguistic measures compared to the Chinese L1 students. There was only one task where no significant difference was found, and this was the rapid naming of digits. Another novel finding is the magnitude of difference between the ENS and the EFL with European L1s. No significant differences were found in these two groups in 6

measures: reading fluency, untimed reading comprehension, total number of words produced in summaries, spelling error rate, phonological awareness, and phonological processing speed. In two of these measures, namely phonological awareness (tested with elision) and reading fluency, the two groups performed almost identically, with only a slight numerical difference. These findings suggest that the large and significant differences found between the group of ENS and the whole group of EFL are driven by Chinese L1 students.

It is worth pointing out that the speed of language processing seems to be an important characteristic describing students' abilities. This is especially evident in two of the measures: reading comprehension and number of words in writing. There was a significant difference between the group of ENS and the group of EFL participants with European L1s in the timed reading comprehension and the writing rate measure. However, when the data were analysed excluding the time factor, the significant differences in these two tasks disappeared, indicating that, if given more time, the EFL with European L1s can obtain scores that do not differ from those of the native speakers of English. The same effect was not observed in the group of EFL with Chinese L1s, and more importantly, the EFL with Chinese L1s were much slower in all the processing tasks when compared to the ENS students. They were two times slower than the ENS when providing answers in the vocabulary task, obtained half of the score in the timed reading comprehension when compared to the ENS, and had a significantly lower writing rate. This may suggest that their speed of language processing is still developing. Another interesting pattern uncovered by the results is the striking similarity in performance of native speakers of English and EFL with European L1s in tasks assessing phonological skills. These tasks included phonological awareness (elision) and rapid automatic naming of digits (RAN digits). No significant differences were found between the ENS and EFL with European L1s in these measures.

The results show that large and significant differences in language and literacy skills at the start of university cannot be generalised to all groups of international students. Despite the fact that the EFL with European L1s obtained slightly greater mean scores than ENS in two measures, the results showed that the ENS are still significantly better than the whole international sample. Since the international students are heterogeneous in terms of their first language, cultural differences, and exposure to English, these factors should be taken into consideration. All three groups were at the same level of cognitive abilities such as fluid

intelligence and working memory, which excludes the possibility that the results are influenced by any of these factors.

Chapter 5: T2 results

This chapter presents results that are based on data collected at Time 2. To collect these data, the students who took part in the study at T1 were asked to repeat the same battery of tasks after one year. Two of the tasks, the matrix reasoning assessing non-verbal intelligence and digit span backward, the assessment of working memory, were administered at T1 only because these abilities were not expected to change (Chierchia et al., 2019; Dobbs & Rule, 1989). The purpose of this procedure was to capture the changes in linguistic knowledge and literacy skills after one year in an English-medium university in order to understand how language develops in this context and whether there are group differences with regards to that development. In addition to that, the three groups of participants were compared at T2 to find out whether the large and significant differences observed at T1 had disappeared and specifically whether the gap observed between the ENS and each of the EFL groups had closed. The T2 results presented in this chapter pertain to answering the second research question:

RQ2. How much do knowledge of English and literacy skills change in the first year at university in home students and in international students speaking Chinese and European L1s? Do international students close the gap on any measures?

This chapter is organised as follows. The participants involved in the T2 data collection are described first. This is because not all participants at T1 returned to T2 testing sessions; the subsample of those who returned at T2 is described again in terms of their demographics. This is followed by a detailed description of the T2 recruitment process. Next, the new mode of data collection necessitated by the onset of the Covid-19 pandemic is presented because it differed from that used at T1. The procedure and instruments used at T2 are described here again as these aspects of the study had to be modified for online delivery for the same reason. It will then be explained how the data were prepared for analysis and analysed. Finally, the results obtained for all 13 measures are presented and followed by a summary of T2 findings.

5.1 T2 participants and recruitment

All the participants who took part in the study at T1 in 2019 were contacted one year after the first data collection. At T2 they were asked to take part in two individual Zoom meetings (rather than in-person meetings as at T1) due to the social distancing rules imposed by the government in response to the Covid-19 pandemic. The T2 recruitment proceeded as follows. The whole sample of students who participated at T1 (*N* = 177) was divided into three subsamples, according to the month of participation. Those who participated in September/October 2019 were contacted again in October 2020, those who participated in November 2019 were approached in November 2020, and finally those who took part in December 2019 and beyond were contacted again in December 2020.

Each participant was contacted up to 4 times, at the beginning of their expected participation month, and at the beginning of each week within this month. The number of invitations sent depended on the speed of their responses to the call for participation. Everyone who received a fourth email was informed that this call was the final one and failing to respond to it would automatically withdraw them from the study. This procedure guaranteed that each participant would repeat all the tasks within approximately one calendar year of the original testing session, with a difference of +/- one month. For example, someone who participated in the T1 study at the beginning of October 2019 and responded to the 4th invitation email at the end of October 2020 would participate within 13 months of the first data collection point. On the other hand, someone who took part in the study at the end of October 2019 and who responded to the first invitation email at the beginning of October 2020 would participate within 11 months of T1. On average, participants took part within 12 months of the first data collection point. Eighty-three per cent of those who participated at T1 returned for the testing sessions at T2.

The whole sample who contributed their data for T2 analysis (*N* = 147), similarly to the sample at T1, consisted of participants who belonged to one of three groups based on their first language characteristics. There were 48 native speakers of English, with 37 (77%) females. The group of EFL with European L1s (*N* = 50) included 36 (72%) females. Participants in this group represented the following language families: Romance (*N* = 30), Germanic (*N* = 7), Hellenic (*N* = 4), Balto-Slavic (*N* = 7), and Finno-Ugric (*N* = 2). The group of

EFL with Chinese L1s (*N* = 49) consisted of speakers of the Mandarin and Cantonese dialects, and included 33 (67%) females.

The unexpected outbreak of Covid-19 in the spring of 2020 made the first year at university an unusual experience. From that time onwards, many students stayed in their home countries abroad for substantial amounts of time during their first year at university. Some others stayed in the UK and did not travel abroad at all. It was therefore important to capture the amount of time the EFL students spent in the UK and outside of the country because this was linked to the amount of exposure to English during the first year at university. The amount of exposure to the target language is a very important factor in second language development and could affect language development trajectories in EFL students. Exposure to English was operationalised in two different ways. The first involved measuring the length of stay outside of the UK, in number of months. The second measure of exposure to English was operationalised in terms of the number of interactions in English and this was measured by the number of friends speaking a particular L1. It was predicted that students who have friends speaking different L1s are exposed to English to a greater extent than those who have friends sharing their first language. Therefore, the EFL students who took part in the study at T2 were asked two additional background questions and compared on both measures. These data helped in understanding both samples and were important when interpreting the main findings of this study.

The EFL students were asked first to state the amount of time spent abroad since the outbreak of the pandemic in mid-March 2020. The data showed that students from the two EFL groups spent a comparable amount of time abroad during their first year at university. It was on average 5 months for both the EFL with European L1s (*M* = 5.02, *SD* = 2.15) and the EFL with Chinese L1s (*M* = 5.0, *SD* = 2.92). The difference in number of months spent abroad was not found to be statistically significant, $t(97) = .14$, $p = .892$. This suggests that students in the two groups were exposed to the second language environment to a similar extent (see Figure 5.1, the error bars represent standard deviations).

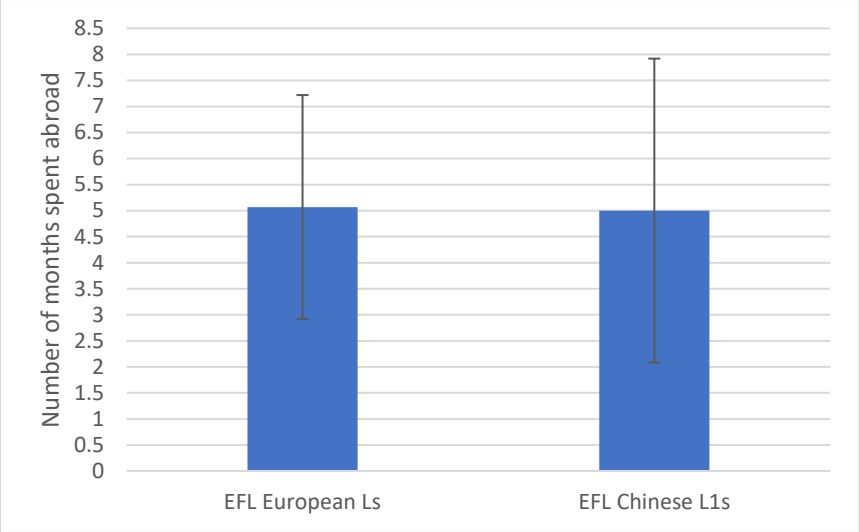

Figure 5. 1 Length of time spent abroad by the EFL students during Year one

To quantify the amount of interaction in English, the EFL students were asked about the first language of the majority of their friends, as an indicator of how frequently they spoke English. They were asked the following question:

During your stay in York last year, the large majority of your friends were:

1. Native speakers of your first language
2. Other international students
3. British home students
4. A balanced mixture of the three
5. Other (please specify)

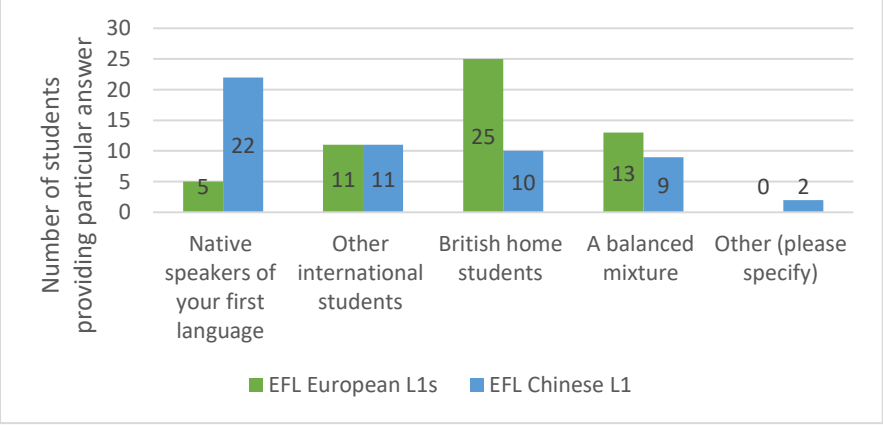

Figure 5. 2 EFL students reporting friends from particular first language categories

Figure 5.2 above shows that a similar number of students in both EFL groups reported having friends who were other international students (20% in each group). Twenty-four per cent (24%) of students speaking European L1s and 16.7% of students speaking Chinese L1 stated that they had a balanced mixture of friends speaking different L1s. The contrast between the two groups becomes apparent in the two categories for friends who are 'Native speakers of your first language' and 'British home students'. Only 9.3% of the European L1 group stated that the majority of their friends were speakers of the same L1, whereas this was true for 40.7% of the Chinese L1 students. An opposite trend exists when it comes to native English-speaking friends. Forty-six per cent (46%) of the European L1 students stated that the majority of their friends were ENS, whereas this was the case for only 18.5% of the Chinese L1 students. This finding that the Chinese L1 students do not have many friends among native speakers of English is consistent with the findings obtained in other studies (Li et al., 2010; Trice, 2003). Taking these proportions together, it can be concluded that the students speaking European L1s used English in daily interaction to a greater extent than the students speaking Chinese L1s. They used English in their daily interactions almost exclusively, whereas the Chinese L1 students seemed to maintain a balance between using English and their first language.

The additional background data obtained from the EFL participants revealed that the two groups of EFL students spent a comparable amount of time outside of the UK. When it comes to the amount of interaction, the data suggest that the L1 European students used English in their day-to-day interaction to a greater extent than the students speaking Chinese L1s.

5.2 T2 testing sessions

According to the original design, the methods and procedure at T2 were supposed to mirror those administered at T1 data collection. However, while the research instruments remained the same as at T1, the mode of data collection changed from face-to-face meetings with the researcher to online Zoom meetings. At T2 each participant attended 2 individual Zoom meetings. Each participant was informed that data collection would be carried out online and was sent a link to the online booking system in the invitation email. They were invited to book two meetings within the same week and instructed that they

would have to stay in a quiet space during both appointments. They were also advised to attend the Zoom session using their laptops or at least iPads, and not phones, a request with which all of them complied.

At the start of both meetings, participants were asked to switch off their mobile phones and not to consult any sources when completing the tasks. In the first meeting, each participant was asked to read and sign an updated consent form sent to them via a link to Qualtrics. They were also informed that they would have to share their screen for some of the tasks. The consent form was followed by the two aforementioned questions about the amount of exposure to English, presented on the screen. In the first meeting, only one task was administered, and similarly to T1 data collection, this was the Vocabulary Size Test (VST), but at T2 it was an individual rather than a group meeting due to constraints imposed by the change in mode of testing.

In the second meeting, participants completed all the remaining tasks: the Nelson-Denny reading test, Which English grammar task, York Adult Assessment-Revised, elision, RAN digits, and sight word efficiency task. The tasks that were computerised at T1 (VST, Nelson-Denny, Which English) were administered via shared links sent using the Zoom chat function. The tasks that were administered in pen-and-paper format at T1 were adapted to suit the Zoom format (they are described below) and the test results were recorded on a new answer sheet prepared for each participant for the T2 testing sessions. Each session lasted between 30 and 60 minutes. At the end of the second meeting, each participant was sent the £15 Amazon voucher and thanked for taking part in the study. Participants were also encouraged to get in touch if they encountered any issues with the vouchers. All Zoom meetings were run by myself. The tasks adapted to the needs of Zoom are described in the next section.

5.3 T2 instruments

Because of the different mode of data collection at T2, pen-and-paper tasks administered in face-to-face meetings at T1 were adapted so that it was possible to administer them on Zoom. Those tasks that were performed on a computer screen at T1 were relatively easy to accommodate in a Zoom meeting. The way in which this was achieved is described below.

Vocabulary Size Test. This task was computerised at T1 and administered in the first group meeting in the computer lab. At T2 the same set of instructions as in T1 was displayed on a shared screen to each participant. Then, they were asked to share their screen and sent a link to the VST software, an access code to the software, and an ID identical to the one they were assigned at T1. In this way it was possible to monitor participants' progress on the task and to ensure they were conducting the task independently and without recourse to outside help. The answers were saved automatically by the software upon task completion.

Nelson-Denny reading comprehension. At T1 this task was computerised and administered via Qualtrics on a computer screen in a face-to-face session. At T2, a link to Qualtrics was shared with participants via the Zoom chat function. The participants had their screens shared with the researcher during task completion to ensure they were adhering to the test procedure and not accessing outside help. The time limit for this task was 15 minutes, after which participants were stopped and asked to save the answers.

Which English grammar test. This was the third and final task that was computerised at T1. Similarly to the procedure for the VST and Nelson-Denny, participants were sent a link on Zoom chat and worked through the answers with their screens shared with the researcher. The answers were saved automatically in Qualtrics upon task completion.

YAA-R. In this task, participants read the same passage as in T1, but at T2 they read it from the screen rather than from paper. Similarly to the T1 instructions, participants were told that they would have to write a summary of the main points once they finished reading. The text was shared with them via Google Sheets with a link to it sent on Zoom chat. At the time the text was accessed, participants were asked whether they were ready to start reading and instructed to signal when they finished. They read silently from their screens shared with the researcher. The time taken to read the passage was recorded on the answer sheets. Participants were then asked to take the pen and paper prepared at the start of the meeting and to write a summary of the main points they recalled from the text. They were given 10 minutes to complete the task. Upon finishing, they were told that they would be asked to take a photo of the text and to send it to the researcher by email at the end of the meeting. This request was postponed until the end of the session, when participants were allowed to use their mobile phones to take the picture. The time for task completion was

recorded on the time sheet in any instance where the task was completed before the time limit.

Elision. In this task, the same set of words as at T1 was presented orally, and participants were asked to repeat each lexical item after the researcher with an indicated sound removed. The participants had a chance to ask for repetition of the target word but audibility on Zoom proved to be reliable and repetitions occurred very rarely. The answers were recorded on the answer sheet.

Rapid Automatic Naming of digits. At T1 the 50-digit stimuli were organised in 10 rows on an A4 sheet of paper. At T2 the digits were copied into a PowerPoint slide in the same layout (50 digits in 10 rows) and presented to participants via the screen share function. The practice items were presented first and they were followed by the test items presented on the next slide. Participants were asked to read the digits from left to right, beginning at the top line. The time taken to perform this task was recorded on the answer sheet.

Sight Word Efficiency. Similarly to RAN, this task was converted from a paper version into a PowerPoint slide and presented to participants via the screen share function. The original layout of lexical items in the paper version (4 columns x 26 words in each column) was modified to suit the PowerPoint format. At T2 the word stimuli were organised into 5 columns (4 columns x 21 words and one column with 20 words), as this proved to be the optimal solution for presenting the words in PowerPoint slides after multiple piloting. The test items were preceded by the practice items.

The exact instructions provided before each task were identical to those used for T1 data collection (see Researcher's Handbook in Appendix F). All T2 tasks were piloted in Zoom meetings with two individuals, one residing in China and one in the UK, to test internet connectivity.

5.4 Time 2 data preparation (outliers, normality, missing values, and task reliability)

The data obtained in computerised tasks were downloaded automatically to Excel; the pen-and-paper tasks were scored, and the results were entered in a spreadsheet manually. The handwritten summaries were typed into a word processor and analysed in the same way as

at T1. In the T2 pool of summaries, 10 words were removed because of lack of legibility. These included three words from across 3 summaries in the ENS group, 5 words from 2 summaries in the EFL with European L1s, and 2 words from 2 summaries in the EFL with Chinese L1. Twenty per cent (20%) of the summaries (10 from each group) were marked by an expert in linguistics to check for scoring consistency. The interclass correlation coefficient based on overall scores was .97, indicating a very high level of agreement between the two raters when scoring summaries independently.

The T2 data comprised two main sets. The first set consisted of the results for each group and measure obtained at T2 (referred to as the T2 main dataset). The second set consisted of the gains obtained in each measure (referred to as gains data), which were calculated by subtracting T1 from T2 scores for each participant and measure.

5.4.1 T2 main dataset

Normality: The T2 data distribution was inspected 1) through normality tests (see Table A.4 in Appendix A for normality test results) and 2) visually through plotting the histograms for each group and measure. According to the Kolmogorov-Smirnoff and Shapiro-Wilk normality tests, not all data sets were normally distributed. All histograms produced for each measure and language group resembled a Gaussian shape, suggesting that the data were normally distributed.

Outliers. The T2 main dataset was inspected for the presence of outliers. The whole data set was split according to language group into three subsets, for ENS, EFL with European L1s, and EFL with Chinese L1s. Each subset was visualised on P-P plots first (created by statistical software); in this way there were three P-P plots for each measure, one for each language group. The observed data points did not diverge significantly from other scores, suggesting the absence of outliers. The data for each language group were also visualised on boxplots, which is a more sensitive method in outlier identification than the eyeball method. The purpose was to check whether any of the outliers detected on boxplots were produced systematically by the same participants or whether a human error was introduced during data handling. The boxplots detected a total of 44 outliers (2.3% of the total data points): this included 43 soft outliers (values between 1.5 and 3 interquartile ranges) and 1 extreme

case (values that are more than three IQRs). A careful inspection showed that none of the ENS students produced more than one outlier. Two participants in the EFL with European L1s sample produced 2 outliers each and three participants in the EFL with Chinese L1s sample produced two outliers each. These findings show that none of the participants produced outlying scores systematically. If they had produced them systematically, it would have suggested that they were careless or fatigued during the testing sessions. For this reason, these outliers were retained in the main analyses, and a median score is reported alongside each mean score.

Missing values: The T2 main dataset contained some missing values. One of the participants from the EFL Chinese L1s group failed to attend the second testing session, leading to a loss of 12 data points (all measures in the second testing sessions). There were two further missing data points, one from an EFL European L1 participant for the spelling error rate (this was caused by a lack of summary readability due to a low-quality picture of the summary). The other missing value was for the RAN measure from an EFL student with Chinese L1 and was caused by human error during data collection. The analyses were conducted with these randomly missing 14 data points excluded. The number of missing values was relatively small and not likely to compromise the power of statistical tests.

Task reliability: The reliability for each task was calculated again at T2 (see Table 5.1). The reliability for the Which English grammar test in the group of ENS was even lower than at T1. As explained earlier, a low result is expected here, and the reason it is lower than at T1 is probably because the ENS students have gained one year of exposure to complex texts while at university and made further improvement in their grammatical knowledge, making it even more difficult to assess this knowledge. The results for the VST, the only task that was standardised on ENS and EFL alike, yielded high reliability scores, but the score is lower for the ENS when compared to T1 results.

Table 5. 1 Time 2 task reliability scores

Task	ENS	EFL European L1s	EFL Chinese L1s
Vocabulary Size Test	.747	.875	.895
Which English grammar test	.502	.687	.778
Nelson-Denny reading comprehension	.872	.870	.875

5.4.2 Gains data

As stated above, the gains for all measures were calculated by subtracting T1 test scores from corresponding T2 values for every participant who took part in the study across two time points. The statistical analyses performed on the raw gains involved descriptive statistics only (standard deviation, mean, median, and range) and therefore the data were inspected for outliers only to check whether there were any errors caused by human error while handling the data.

5.5 Results

The statistical analyses were performed using the SPSS statistical package, version 28.0.0.0. To answer RQ2, mixed-design ANOVAs were performed first to investigate the change in performance across the two time points and whether any of the groups had closed the gap by the time of T2 testing. The mixed-design ANOVAs explained very little in terms of time by group interaction, with non-significant results for almost all measures. However, the post hoc tests on the main effect of group indicated that one of the EFL groups managed to close the gap with the ENS, as the between-group differences for many of the measures were not significantly different. For this reason, *t*-tests were performed for each language group to investigate the magnitude of change across the two time points and Cohen's d was obtained to quantify the amount of change for all three groups. The results of the post hoc tests for data obtained at T2 only was looked at more closely to obtain a full picture of the results.

The results are presented in the following way. The descriptive statistics on the gains made across the two time points (mean, standard deviation, median, and range) come first to present the overall picture of the results. They are followed by the *t*-test and effect size results for within-group differences. The mixed-design ANOVA comes next with the main effect of time and group. The main effect of time shows the difference in performance of all participants at T1 and T2, while the main effect of group shows the results for each group irrespective of the time at which a test was taken. This is followed by a post hoc test to see whether the cumulative scores obtained by each group at both time points are different from each other. Finally, the interaction between time and group is reported. If the interaction was significant (which was the case for one measure only), the gains across all three groups were compared using a one-way ANOVA. See Figure 5.3 below for an outline of the results presentation.

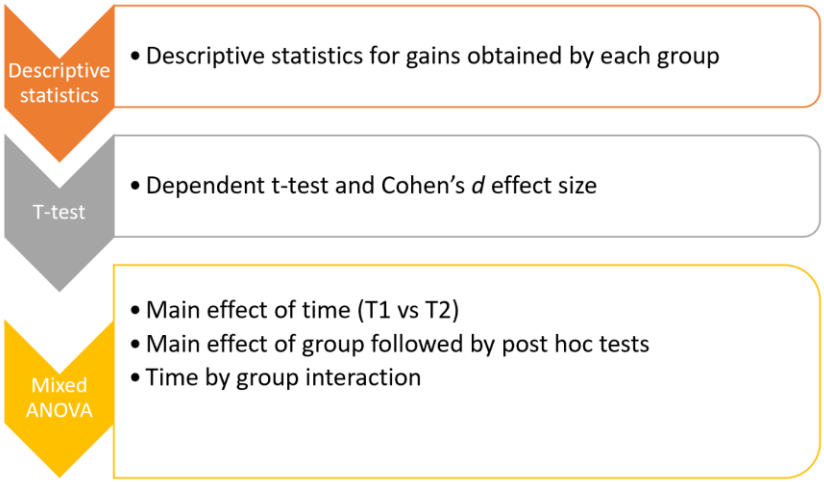

Figure 5. 3 Flow chart representing T2 results presentation

5.5.1 Vocabulary knowledge

The score for the vocabulary test was expressed on a scale between 0 and 140. As can be seen in Figure 5.4, the ENS students had the largest vocabulary size both at T1 and a year later. They were followed by the EFL students with European L1s. The students with Chinese L1s had the smallest vocabulary size at both time points. The groups of ENS and students with European L1s were closer to each other in performance than the two EFL groups were to each other at both time points. All three groups improved their scores. The largest change in the vocabulary test was detected in the EFL European L1s' group, rising from 104.55 (N = 49, SD = 11.06) at T1 to 108.82 (N = 49, SD = 11.30) at T2, with a difference of

4.27 (*N* = 49, *SD* = 5.24, *Mdn* = 4, Range = -9–20). This increase was statistically significant with a large effect size, *t*(48) = -5.70, *p* < .001, Cohen's *d* = -.815. The Chinse L1 group made an average gain of 3.86 (*N* = 49, *SD* = 7.92, *Mdn* = 4, Range = -16–20), improving the vocabulary score from 73.69 (*N* = 49, *SD* = 14.70) at T1 to 77.55 (*N* = 49, *SD* = 14.96) at T2. This change was also statistically significant, with a medium effect size as indicated by the *t*-test, *t*(48) = -3.41, *p* < .001, Cohen's *d* = -.487. The smallest improvement was observed in the ENS group, who improved their score by 1.71 (*N* = 48, *SD* = 4.28, Mdn = 2, Range = -9–9), from 118.56 at T1 (*N* = 48, *SD* = 7.79) to 120.27 (*N* = 48, *SD* = 6.40) at T2. This change was also statistically significant, with a small to medium effect size, *t*(47) = -2.76, *p* = .004, Cohen's *d* = -.399. The error bars in all the following figures represent standard errors.

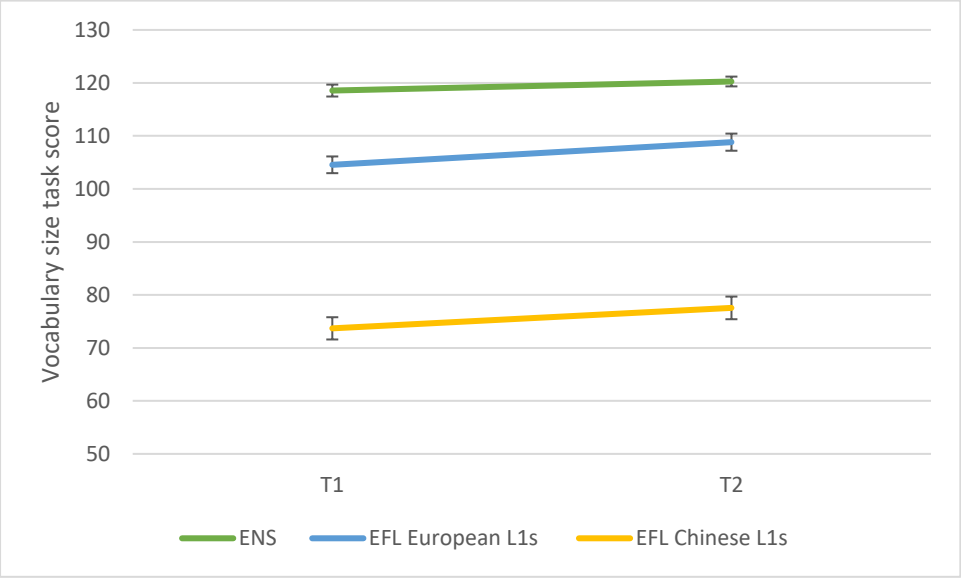

Figure 5. 4 VST score change across two time points in all groups

The mixed-design ANOVA confirmed that the main effect of time was significant, *F*(1, 143) = 43.18, *p* < .001, *r* = 0.48, which means that there were significant changes over time in the vocabulary test irrespective of the first language spoken. In general, the whole sample performed better at T2 (*N* = 146, *M* = 102.09, *SD* = 21.40) than at T1 (*N* = 146, *M* = 98.80, *SD* = 22.03). The main effect of group was also significant, *F*(2, 143) = 200.06, *p* < .001, *r* = 0.76. This means that some groups obtained greater cumulative mean scores across both time points than others. A further post hoc test revealed a significant difference between the group of ENS (*M* = 119.42) and each of the EFL groups: the one with European L1s (*M* = 106.68), *p* =.000, and the one with Chinese L1s (*M* = 75.62), *p* < .001. A significant difference

was also found between the two EFL groups of students ($p < .001$). No significant interaction between time and group was found, $F(2, 143) = 2.51$, $p = .09$, $r = .13$. In sum, all three groups improved their vocabulary knowledge across the two time points and all three groups did so significantly. The group that improved to the greatest extent was the group of EFL with European L1s. Despite the large improvement, the difference between them and the ENS students remains statistically significant at T2.

In addition to the raw scores obtained in the vocabulary task, the estimated vocabulary sizes calculated by the VST algorithm are also presented here to allow some direct comparisons between the results obtained here and those from other studies. According to the VST algorithm calculated on raw scores, all three groups had higher estimated vocabulary size at T2 than at T1 and the group of EFL with European L1s made somewhat larger gains in comparison to the other two groups of students. The group of EFL with European L1s gained over 1,000 word families (*N* = 49, *M* = 1,263, *SD* = 1,571.05, Mdn = 800, Range = 1,000–5,100). The group of ENS gained on average over 700 word families (*N* = 48, *M* = 769, *SD* = 2,196.72, Mdn = 1,000, Range = -4,900–4,600). The EFL with Chinese L1s' gain was substantially lower when compared to the other two groups; they showed an average increase of over 400 word families after one year at university (*N* = 49, *M* = 410, *SD* = 938.80, Range = -2,000–3,300). Across both time points, the ENS group performed better on this task than the two groups of EFL students.

5.5.2 Vocabulary test response times

As can be seen in Figure 5.5, the ENS students were the fastest in providing responses in the vocabulary test at both time points. They were followed by the EFL students with European L1s. Students with Chinese L1s were the slowest at both time points. However, the groups of ENS and students with European L1s were closer to each other in performance than the two EFL groups were to each other. All three groups improved their speed significantly and the group that made the greatest improvement, as indicated by the effect size, was the group of EFL with European L1s. Their speed improved on average by 727 milliseconds across the two time points (*N* = 49, *M* = -727, *SD* = 1,172, *Mdn* = -534, Range = -3,424–2,189), dropping from 6,744 at T1 to 6,017 at Time 2, with a medium to large effect size, $t(48) = 4.34$, $p < .001$, Cohen's $d = .620$. The second most improved group was the ENS

students, whose reaction time dropped from 5,119 at T1 to 4,599 milliseconds on average at T2, a difference of 519 milliseconds (N = 48, M = -519, SD = 871, Mdn = -523.5, Range = -2213–2197). The effect size was also medium to large, $t(47) = 4.13$, $p = .00$, Cohen's $d = .597$. The group of EFL with Chinese L1s experienced a drop from 8,959 at T1 to 8,046 at T2, an improvement of 912 milliseconds (N = 49, M = -912, SD = 1,852, Range = -6,438–3,273), with a medium effect size, $t(48) = 3.45$, $p < .001$, $d = .493$.

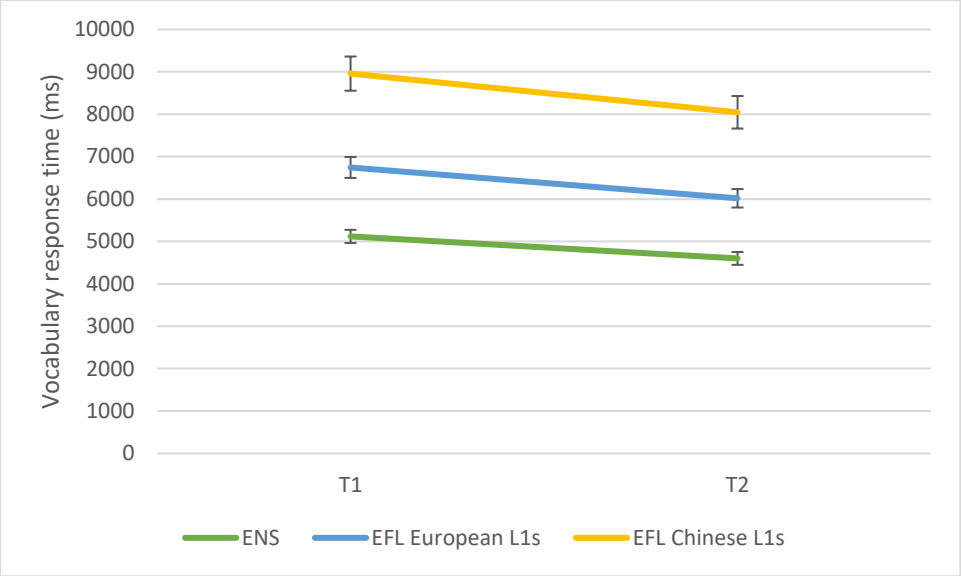

Figure 5. 5 VST response time change across two time points in all three groups

The mixed-design ANOVA confirmed that there was a significant main effect of time, $F(1, 143) = 40.60$, $p < .001$, $r = .47$, which means that there are significant changes in the speed of providing correct answers in the Vocabulary Size Test across the two time points. On average, the mean time for all three groups decreased from T1 (N = 146, M = 6,953, SD = 2,546) to T2 (N = 146, M = 6,232, SD = 2,349), indicating an improvement in the speed for providing correct answers across the two time points. There was also a significant main effect of group, $F(2, 143) = 48.70$, $p < .001$, $r = .96$. A further post hoc test revealed a significant difference between the ENS (M = 4,859) and both EFL groups: students with European L1s (M = 6,381), p =.000, and students with Chinese L1s (M = 8,502), $p < .001$. The difference between the two EFL groups was also significant, $p < .001$. No significant interaction was found between time and group, $F(2, 143) = 1.00$, $p = .369$, $r = 0.08$. In sum, all three groups improved their speed across the two time points, as shown by the t-tests

and further confirmed by the main effect of time in the mixed-design ANOVA. The group that improved the most, as indicated by the effect size, was the group of EFL with European L1s, but this group did not manage to close the gap that existed between them and the group of ENS at T1. This was indicated by the between group differences in the mixed-design ANOVA and further it was further confirmed by the results obtained on T2 dataset.

5.5.3 Grammar

In the Which English grammar test, it was possible to obtain a maximum of 95 points. As can be seen in Figure 5.6, all three groups improved their scores in the grammar test; however, the group of ENS performed best as indicated by their mean scores obtained at both time points. The group of EFL with European L1s was the second best in this task and they were much closer in their performance to the ENS than to the other EFL group. The group that improved the most was the EFL with European L1 students, with an increase in their mean score from 89.76 (*SD* = 3.93) at T1 to 90.58 (*SD* = 3.48) at T2, an increase of .82 (*N* = 50, *M* = 0.82, *SD* = 3.29, *Mdn* = 1, Range = -5–7). This change proved to be statistically significant with a small effect size, *t*(49) = -1.76, *p* = .042, Cohen's *d* = -.249. The other two groups did not improve significantly. The group of EFL with Chinese L1s improved their mean score from 80.52 (*SD* = 6.66) at T1 to 81.29 (*SD* = 6.46) at T2, an improvement of 0.78 (*N* = 48, *M* = 0.78, *SD* = 5.51, *Mdn* = 1, Range = -14–15), *t*(47) = -.97, *p* = .169, Cohen's *d* = -.140. The group of ENS improved by only .06 across the two time points (*N* = 48, *M* = 0.06, *SD* = 2.72, *Mdn* = 0, Range = -5–7), which was an increase from 91.44 (*SD* = 2.80) to 91.50 (*SD* = 2.48), but again this improvement was not statistically significant, *t*(47) = -.16, *p* = .437, *d* = -.023.

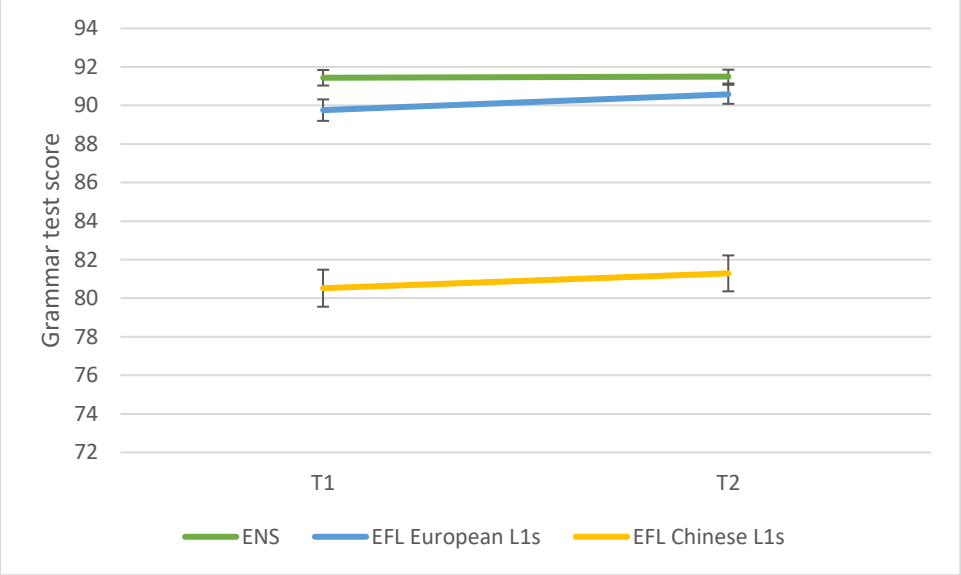

Figure 5. 6 Grammar test score change across two time points in all three groups

According to a mixed-design ANOVA, there was no significant main effect of time, $F(1, 143) = 2.75$, $p = .099$, $r = 0.14$, which means that there are no significant changes across the whole sample in terms of grammar development; this is consistent with what the *t*-tests show. The average number of correct answers at both time points in all participants remained very stable from T1 ($N = 146$, $M = 87.27$, $SD = 6.715$) to T2 ($N = 146$, $M = 87.83$, $SD = 6.390$). There was a significant main effect of group, $F(2, 143) = 93.347$, $p < .001$, $r = 0.62$, which indicates that some groups obtained higher cumulative scores at both time points. A further post hoc test revealed a lack of significant difference between the group of ENS ($M = 91.47$) and the group of EFL with European L1s ($M = 90.17$), $p = .067$. A significant difference was found between the group of ENS and EFL with Chinese L1s ($M = 80.91$), $p < .001$. A significant difference was also found between the two EFL groups, $p < .001$. No significant interaction between time and group was found, $F(2, 143) = .540$, $p = .584$, $r = 0.06$. These results showed that only the group of EFL with European L1s recorded a significant positive change in the grammar task and, despite the effect being only small, they were able to catch up with the group of ENS. This was confirmed by a lack of significant difference a post hoc test following the main effect of group in a mixed-designed ANOVA, and in another post hoc test run on T2 data.

5.5.4 Reading fluency

The test result in the reading fluency task constituted the number of words read correctly within 45 seconds from a list of 104 English words. It can be seen from Figure 5.7 that all three groups read more of the words presented in the task at T2 than at T1, with almost identical performance in the ENS and EFL with European L1s at both time points. All three groups read significantly more words at T2, with the group of EFL with Chinese L1s making the largest improvement. They gained 2.38 words on average (N = 48, M = 2.38, SD = 5.61, Mdn = 2, Range = -7–21), which is a difference between the mean score of 78.10 (SD = 9.16) obtained at T1 and 80.48 (SD = 9.92) obtained at T2. The effect size proved to be medium as indicated by the t-test, $t(47) = -2.93$, $p = .003$, Cohen's $d = -.423$. The group of ENS students improved their score from 90.46 (SD = 9.24) at T1 to 91.96 (SD = 9.27) at T2. This was an improvement of 1.50 (N = 48, M = 1.50, SD = 6.10, Mdn = 0.5, Range = -14–22). The effect size was small, $t(47) = -1.70$, $p = .047$, Cohen's $d = -.246$. The group of EFL with European L1s improved to a similar extent to the ENS. They obtained a mean score of 90.56 (SD = 8.68) at T1 and 91.76 (SD = 8.04) at T2, with a change of 1.20 (N = 50, M = 1.20, SD = 5.01, Mdn = 1, Range = -15–10). The effect size was also small, $t(49) = -1.70$, $p = .048$, $d = -.240$.

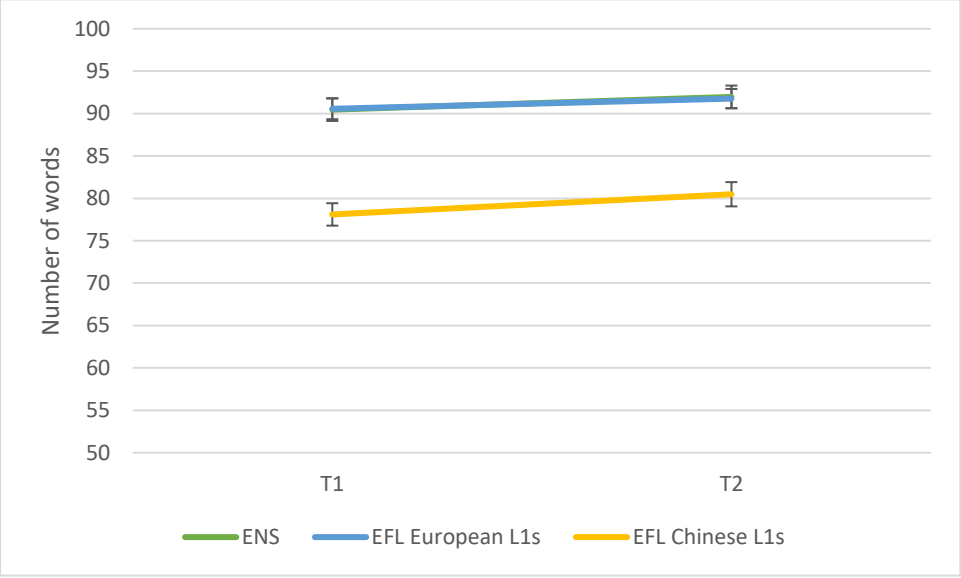

Figure 5. 7 Sight word efficiency test score change across two time points

The main effect of time in a mixed-design ANOVA proved to be significant, $F(1, 143) = 13.406$, $p < .001$, $r = 0.29$. On average, the number of words produced by the whole sample increased from T1 (N = 146, M = 86.43, SD = 10.70) to T2 (N = 146, M = 88.12, SD = 10.51).

There was a significant main effect of group, $F(2, 143) = 30.651$, $p < .001$, $r = 0.42$. A further post hoc test revealed a lack of significant difference between the group of ENS ($M = 91.21$) and students with European L1s ($M = 91.16$), $p = .999$. There was a significant difference between the group of ENS and EFL with Chinese L1s ($M = 79.29$), $p < .001$, and a significant difference was found between the two EFL groups of students, $p < .001$. No significant interaction between time and group was found, $F(2, 143) = .582$, $p = .560$, $r = 0.06$. In sum, all three groups improved significantly across the two time points, which is further confirmed by the significant main effect of time in the mixed ANOVA. The group that improved the most was the EFL with Chinese L1s, but the medium effect size was not sufficient to close the distance between this group and the other two groups of students. This is indicated by significant between group differences in a post hoc tests in the mixed-design ANOVA as well as significant group differences at T2.

5.5.5 Reading rate

The reading rate was expressed in terms of the number of words read per minute during a silent passage reading task. Figure 5.8 shows that not all groups improved in this task, which is contrary to the prediction. Only one group recorded an increase in the number of words read per minute and this was the EFL with European L1s. However, it was the group of ENS who, despite a negative trend in performance with a decrease from T1 to T2, obtained the greatest mean scores at each time point. The group of EFL with European L1s read on average 5.68 more words across the two time points ($N = 50$, $M = 5.68$, $SD = 47.57$, $Mdn = -1.80$, Range = -113.40–135.12). This is a result of difference between 174.18 ($SD = 67.95$) words read at T1 and 179.86 ($SD = 81.99$) at T2. This change, however, was not statistically significant, $t(49) = -.84$, $p = .201$, Cohen's $d = -.119$. The group of ENS recorded a decrease of 20.31 words read per minute on average ($N = 48$, $M = -20.31$, $SD = 56.12$, $Mdn = -11.07$, Range = -187.4 –102.85), as they obtained a mean score of 218.10 ($SD = 77.71$) at T1 and a lower mean score of 197.79 ($SD = 68.25$) at T2. This change was statistically significant with a small effect size, $t(47), = 2.51$, $p = .008$, $d = .362$. The group of EFL with Chinese L1s also experienced a drop in their performance, as they obtained a mean score of 135.07 ($SD =$ 53.08) at T1 and 127.22 ($SD = 56.93$) at T2, making a loss of 7.86 words read per minute on

average (*N* = 48, *M* = -7.86, *SD* = 57.81, *Mdn* = -11.36, Range = -143.21–181.77). This change, however, was not statistically significant, *t*(47) = .94, *p* = .176, Cohen's *d* = 136.

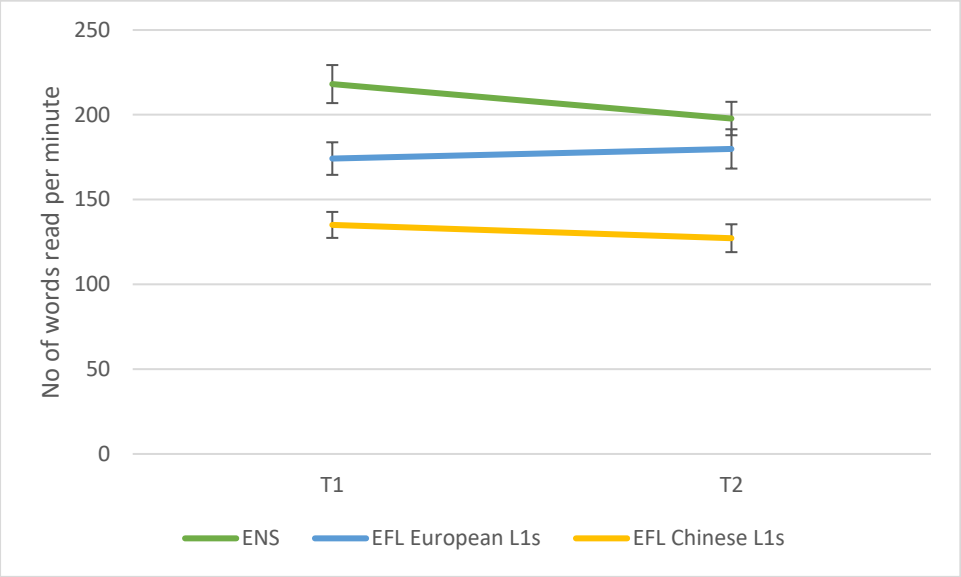

Figure 5. 8 Reading rate change across two time points

On average, the number of words read per minute decreased from T1 (*N* = 146, *M* = 175.76, *SD* = 74.66) to T2 (*N* = 146, *M* = 168.45, *SD* = 75.66), which is contrary to the prediction that the reading rates would improve, alongside other skills. However, this change was not statistically significant as confirmed by the mixed-design ANOVA, *F*(1, 143) = 2.82, *p* = .095, *r* = 0.14. A significant main effect of group was found, *F*(2, 143) = 18.06, *p* < .001, *r* = 0.24. A further post hoc test revealed a significant difference between the groups of ENS (*M* = 207.95) and EFL with European L1s (*M* = 177.02), *p* = .049. There was a significant difference between the groups of ENS and EFL with Chinese L1s (*M* = 131.15), *p* < .001. The difference between the two EFL groups was also significant (*p* = .001). No significant interaction between time and group was found, *F*(2, 143) = 2.847, *p* = .610, *r* = 0.14. Despite the significant difference between the group of ENS and EFL with European L1s revealed by the post hoc test in the mixed ANOVA, the EFL with European L1s were not distinguishable from the ENS at T2; this, however, is due to the significant drop in the performance of the ENS students rather than to the EFL with European L1s catching up with the ENS students.

5.5.6 Timed reading comprehension

Nelson-Denny timed reading comprehension was scored on a scale between 0 and 38. In general, all three groups performed better at T2 than at T1 in the Nelson-Denny reading comprehension task. As can be seen in Figure 5.9, the group of ENS was the one that obtained the highest mean reading comprehension scores at the two time points, and the group with Chinese L1s obtained the lowest mean scores. The groups of ENS students and students with European L1s were closer to each other in performance than the two EFL groups were to each other.

The group that improved the most and significantly in this task was the EFL with European L1s, who improved on average by 1.68 (*N* = 50, *M* = 1.68, *SD* = 4.83, *Mdn* = 1.5, Range = -12–13), with a mean score of 20.76 (*SD* = 6.79) obtained at T1 and 22.44 (*SD* = 6.79) at T2. This difference proved to be statistically significant with a small to medium effect size, *t*(49) = -2.46, *p* = .009, Cohen's *d* = -.348. The other EFL group was the second-best group in the amount of improvement. They gained on average 0.94 (*N* = 48, *M* = .94, *SD* = 4.04, *Mdn* = 0, Range = -9–8), improving from 12.33 (*SD* = 6.10) at T1 to 13.27 (*SD* = 6.47) at T2, but this change was not statistically significant as indicated by the *t*-test, *t*(47) = -1.61, *p* = .057, Cohen's *d* = -.232. The group of ENS improved the least, with an average gain of only 0.27 (*N* = 48, *M* = 0.27, *SD* = 4.79, *Mdn* = 0, Range = -10–15), with mean comprehension scores of 25.29 (*SD* = 5.95) at T1 and 25.56 (*SD* = 6.75) at T2. This gain also was not statistically significant, *t*(47) = -.39, *p* = .348, Cohen's *d* = -.057.

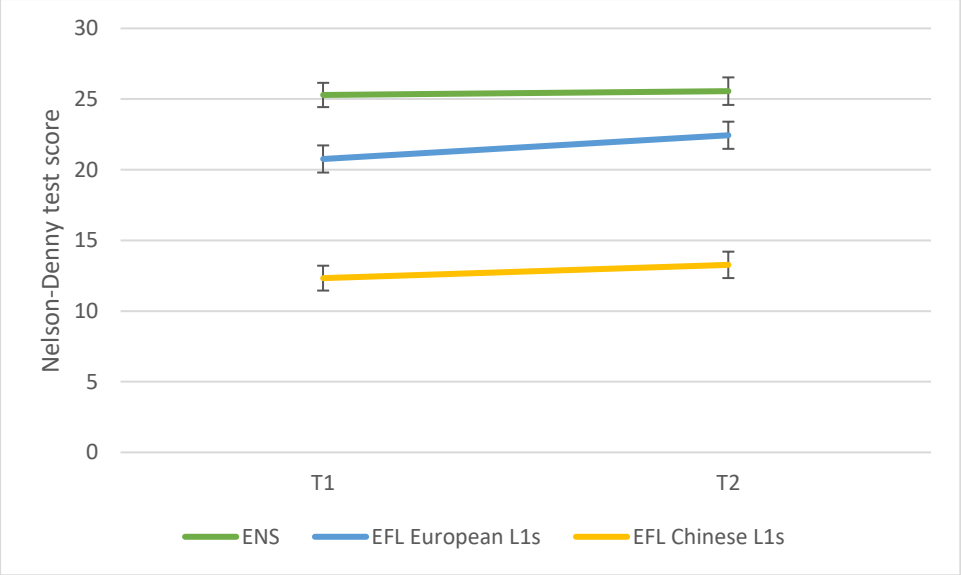

Figure 5. 9 Nelson-Denny timed reading comprehension change across two time points

The results of a mixed-design ANOVA confirmed a significant main effect of time, $F(1, 143) = 6.481$, $p = .012$, $r = 0.21$. On average, the number of correct answers provided in the reading comprehension test increased from T1 ($N = 146$, $M = 19.48$, $SD = 8.23$) to T2 ($N = 146$, $M = 20.45$, $SD = 8.43$). The main effect of group was also significant, $F(2, 143) = 54.63$, $p < .001$, $r = 0.53$. This indicates that there were between-group differences in cumulative scores between all three groups. A further post hoc test revealed a significant difference between the group of ENS ($M = 25.43$) and both EFL groups of students, the one with European L1s ($M = 21.60$), $p = .007$, and the one with Chinese L1s ($M = 12.80$), $p < .001$. The difference between the two EFL groups of students was also significant, $p < .001$. No significant interaction between time and group was found, $F(2, 143) = 1.167$, $p = .314$, $r = 0.09$. The overall cumulative scores for the group of ENS and the group of EFL with European L1s show significant differences between these two groups. However, a further post hoc test performed on T2 data showed that the EFL speaking European L1s managed to catch up with the ENS by the time they started their second year at university as no significant difference between these two groups existed in their T2 scores.

5.5.7 Untimed reading comprehension

The ratio of correct to total attempted answers in the Nelson-Denny task was also analysed; this constitutes the untimed reading comprehension score. Figure 5.10 shows that all three

groups obtained a lower mean score at T2, which is contrary to the predicted positive change. The group that performed best at both time points was the group of ENS, and the EFL with Chinese L1s obtained the lowest scores at both time points. The group of EFL with European L1s is placed in the middle but these students were closer in their performance to the ENS students than to the other EFL group. The group of ENS obtained a mean score of 0.83 at T1 (*SD* = .10) and .82 at T2, with a loss of .01 across the two time points (*N* = 48, *SD* = 0.11, *Mdn* = -.009, Range = -0.308–.268). This difference was not statistically significant, *t*(47) = .67, *p* = .253, Cohen's *d* = .097. The group of EFL with European L1s obtained a mean score of 0.79 at T1 (*SD* = .12) and 0.78 (*SD* =.13) at T2, making a loss of .01 (*N* = 50, *SD* = .132, *Mdn* = -.001, Range = -.473–.302). This change also was not statistically significant, *t*(49) = .42, *p* = .340, Cohen's *d* = .059. The group of EFL with Chinese L1s obtained a mean score of 0.70 (SD = .14) at T1 and 0.67 (SD = .19) at T2, with a loss of .03 on average (*N* = 48, *SD* = .17, *Mdn* = -.020, Range = -.397–.384). As for the other two groups, this change was not statistically significant, *t*(47), *p* = .129, Cohen's *d* = .165.

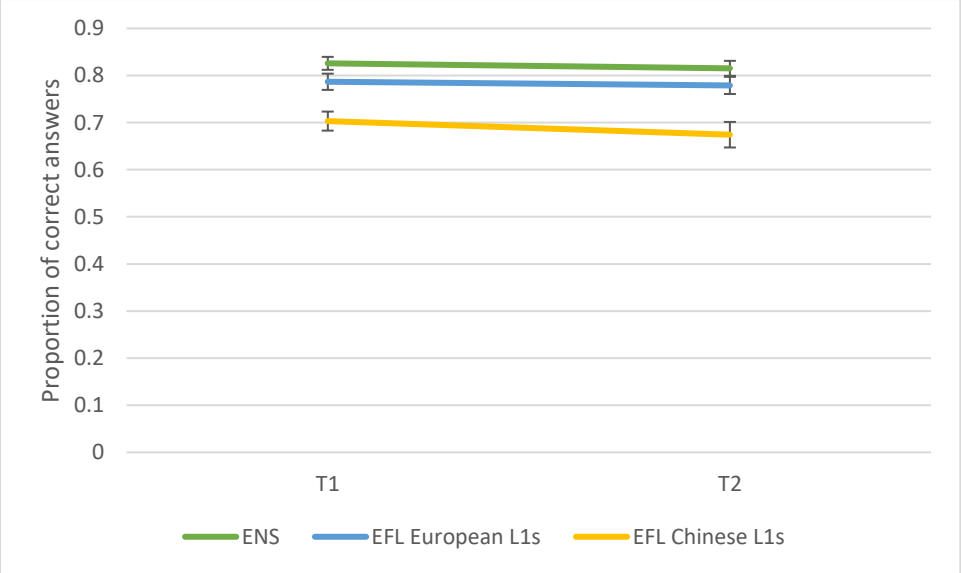

Figure 5. 10 Nelson-Denny untimed comprehension score change across two time points

According to a mixed-design ANOVA, the main effect of time was not significant, *F*(1, 143) = 1.81, *p* = .181, *r* = 0.11. The reading rates decreased slightly from T1 (*N* = 146, *M* = .77, *SD* = .13) to T2 (*N* = 146, *M* = .76, *SD* = .16), which is consistent with the results obtained in the *t*-tests. The main effect of group was significant, *F*(2, 143) = 17.12, *p* < .001, *r* = 0.33. The post

hoc test revealed that there was no significant difference between the ENS and the group with European L1s (p = .140), there was a significant difference between the ENS and EFL with Chinese L1s, p < .001, and there was a significant difference between the two EFL groups of students, p < .001. The time by group interaction was not significant, $F(2, 143)$ = .32, p = .727, r = 0.05. In sum, all three groups performed less well across the two time points in this task. The initial lack of significant difference between the group of ENS and EFL with European L1s holds across the two time points, as indicated by a post hoc on the mixed-design ANOVA and further confirmed by a post hoc test on T2 results.

5.5.8 Total number of words

In the writing task, the participants were asked to write a summary of the main points they could recall after reading the passage *The History of Chocolate*. Four measures were obtained in this task: summary length expressed as the total number of words produced, writing rate, spelling error rate, and summarisation skills expressed as number of content points recalled. The results for each of these measures will be presented in turn.

In the first measure quantifying writing skills, summary length, two of the groups produced fewer words on average in their summaries at T2 than at T1, which is contrary to the prediction that the scores would improve across the two time points. Figure 5.11 shows that only one group improved its result and this was the EFL with European L1s. Despite this pattern of performance, the group that obtained the highest mean scores at both time points was the group of ENS; the group of EFL with European L1s was closer in their performance to this group than to the other group of EFL students. The group of EFL with European L1s improved by 5 words (3.4%) on average (N = 50, M = 5.06, SD = 35.26, Mdn = 5, Range = -73–81). They obtained mean scores of 145.40 (SD = 40.49) at T1 and 150.46 (SD = 40.46) at T2. This gain, however, was not significant as indicated by the t-test, $t(49)$ = -1.02, p = .158, Cohen's d =-.143. The group of ENS produced on average 5 words (3%) less across T1 (M = 164.56, SD = 34.47) and T2 (M = 159.48, SD = 38.88), with a mean difference of 5.08 between these scores (N = 48, M = -5.08, SD = 30.34, Mdn = -7.5, Range = -66–57). This drop, however, was not statistically significant, $t(47)$ = 1.16, p = .126, Cohen's d = .168. The group of EFL with Chinese L1s also produced fewer words at T2; their score dropped by 1 word on average (1%) (N = 48, M = -1.27, SD = 30.82, Mdn = 5.5, Range = -80–65). This

represents the difference between mean scores of 117.71 (*SD* = 30.26) obtained at T1 and 116.44 (*SD* = 37.71) at T2. This difference, however, was not statistically significant, *t*(47) = .29, *p* = .388, Cohen's *d* = .041.

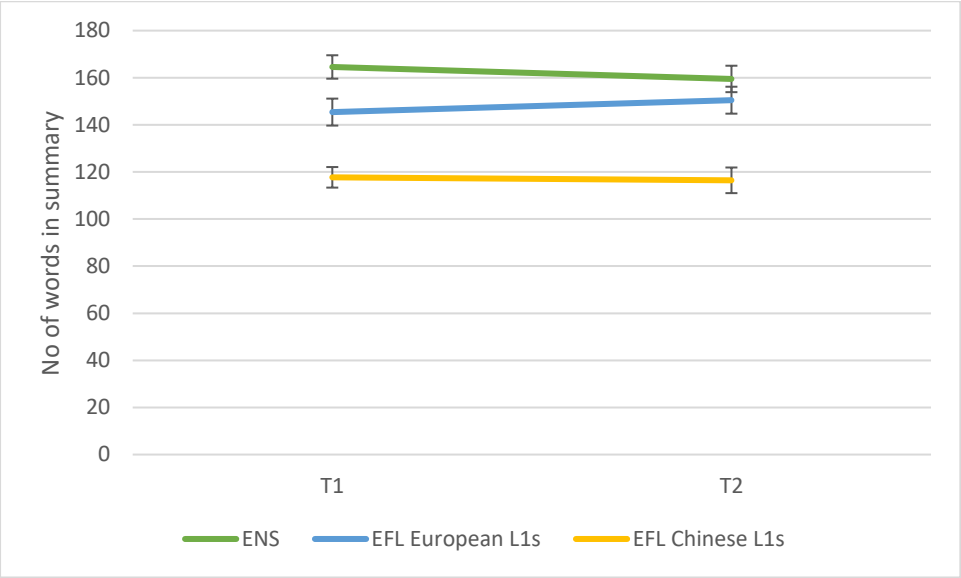

Figure 5. 11 Number of words produced in writing task across two time points

According to a mixed-design ANOVA, there was no significant main effect of time, *F*(1, 143) = .03, *p* = .872, *r* = 0.01. The average number of words produced in the writing task was almost identical at T1 (*N* = 146, *M* = 142.60, *SD* = 40.04) and T2 (*N* = 146, *M* = 142.24, *SD* = 42.97). There was a significant main effect of group, *F*(2, 143) = 22.498, *p* < .001, *r* = 0.37. A further post hoc test revealed that there was no significant difference between the group of ENS (*M* = 162.02) and the EFL with European L1s (*M* = 147.93), *p* = .119, but a significant difference was found between the ENS and the EFL with Chinese L1s (*M* = 117.07), *p* < .001, and another significant difference between the two EFL groups, *p* < .001. There was no significant interaction between time and group, *F*(2, 143) = 1.239, *p* = .293, *r* = 0.09. The initial difference that existed between the groups of EFL and ENS disappeared with time, as indicated by the lack of significant difference in a post hoc test in a mixed design ANOVA and another post hoc test performed on T2 data. This effect, however, is driven by the drop in mean scores in the ENS group.

5.5.9 Writing rate

The writing rate was expressed as number of words written per minute in the summarisation task. As can be seen in Figure 5.12, all three groups improved their writing rate across the two time points. The ENS were those who obtained the greatest mean scores at both time points, the EFL with European L1s performed very similarly to this group at both time points, and the EFL with Chinese L1s obtained the lowest mean scores. The group that improved to the greatest extent was the EFL with European L1s. They improved on average by almost two words written per minute (*N* = 50, *M* = 1.83, *SD* = 3.52, *Mdn* = 2.25, Range = -5.6–8.39). Their mean score obtained was 16.56 (*SD* = 3.98) at T1 and 18.39 (*SD* = 3.55) at T2, and they were the only group who improved significantly, with medium effect size as indicated by the *t*-test, *t*(49) = -3.68, *p* < .001, Cohen's *d* = -.521. The group of ENS obtained a mean score of 18.72 (*SD* = 3.76) at T1 and 19.31 (*SD* = 3.70) at T2, which is an improvement of about half a word on average (*N* = 48, *M* = 0.59, *SD* = 3.26, *Mdn* = 0.58, Range = -7.85–9.72); this change, however, was not statistically significant, *t*(47) = -1.25, *p* = .109, Cohen's *d* = -.180. The group of EFL with Chinese L1s' gain was about half a word on average (*N* = 48, *M* = 0.55, *SD* = 3.44, *Mdn* = 1.12, Range = -10.30–6.66). They obtained a mean score of 13.61 (*SD* = 3.36) at T1 and 14.16 (*SD* = 3.97) at T2, but this gain was not statistically significant, *t*(47) = -1.10, *p* =138, Cohen's *d* = -.159.

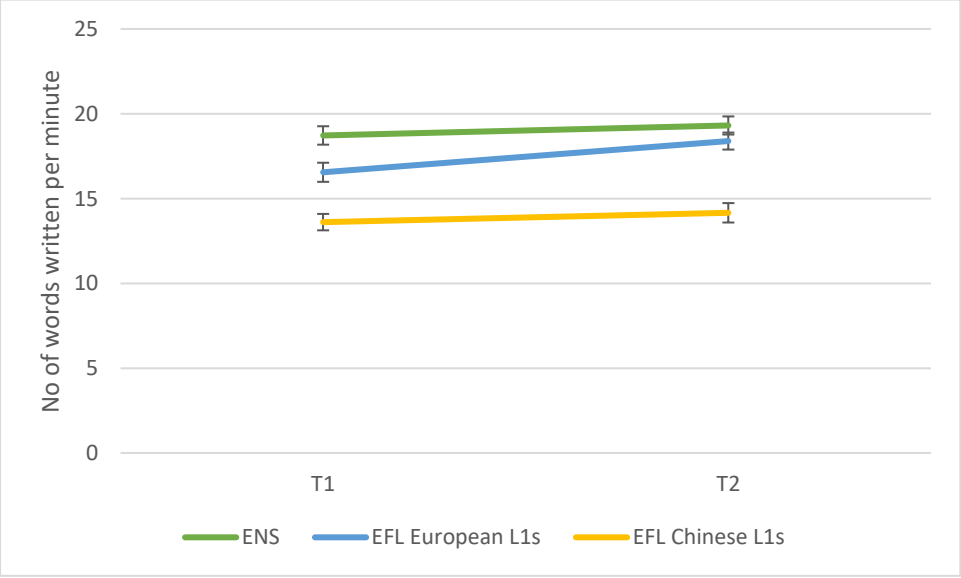

Figure 5. 12 Writing rate change across two time points

A mixed-design ANOVA confirmed a significant main effect of time, $F(1, 143) = 12.29$, $p < .001$, $r = 0.28$. This means that there was significant improvement for all three groups taken together over one year from T1 ($N = 146$, $M = 16.30$, $SD = 4.24$) to T2 ($N = 146$, $M = 17.30$, $SD = 4.34$). The main effect of group was also significant, $F(2, 143) = 30.36$, $p < .001$, $r = 0.42$, which suggests that some of the groups had higher cumulative scores higher than others. A post hoc test revealed that there was no significant difference between the ENS ($N = 19.02$) and the EFL with European L1s (M = 17.47), $p = .063$, but a significant difference was found between the ENS and the EFL with Chinese L1s (M = 13.88), $p < .001$, and a significant difference was found between the two EFL groups, $p < .001$. The time by group interaction was not significant, $F(2, 143) = 2.27$, $p = .107$, $r = 0.13$. In sum, all three groups improved their writing rate but only the group of EFL with European L1s recorded a statistically significant improvement across the two time points. This improvement allowed them to catch up with the ENS, as the difference between these two groups of students disappeared with time. This is indicated by the post hoc test following the main effect of group in mixed design ANOVA, and another post hoc test based on T2 data only.

5.5.10 Summarisation skills

The scores for summarisation skills were based on the number of relevant content points recalled in the summary, with a maximum of 20 points available. In general, all three groups increased their mean scores in this task across the two time points (see Figure 5.13). The group of ENS students performed better than the two groups of EFL students, and the group of EFL with Chinese L1s obtained the lowest mean scores. The group of EFL with European L1s were in between those two groups but closer in their performance to the ENS students. The group of EFL with Chinese L1s improved the most from among all three language groups as indicated by the effect size, with an increase of their mean score from 4.31 ($SD = 2.10$) at T1 to 5.08 (SD = 2.44) at T2, a difference of over half a point on average ($N = 48$, $M = 0.77$, $SD = 1.99$, $Mdn = .05$, Range = -2–6). This change proved to be statistically significant, $t(47) = -2.68$, $p = .005$, Cohen's $d = -.387$. The second most improved group was the EFL with European L1s. This group obtained a mean score of 6.90 ($SD = 2.68$) at T1 and 7.70 ($SD = 3.11$) at T2, making a gain of almost one point on average ($N = 50$, $M = 0.80$, $SD = 3.10$, $Mdn = 0$, Range = -4–12). This change reached statistical significance with a small effect size as

indicated by the *t*-test, *t*(49) = -1.83, *p* = .037, Cohen's *d* = -.258. The group of ENS obtained a mean score of 8.31 (*SD* = 2.68) at T1 and 8.79 (*SD* = 2.45) at T2. They improved on average by around half a point (*N* = 48, *M* = 0.48, *SD* = 2.61, *Mdn* = 0, Range = -6–6); their gain was the smallest from among all three groups and not statistically significant, *t*(47) = -1.27, *p* = .105, Cohen's *d* = -.184.

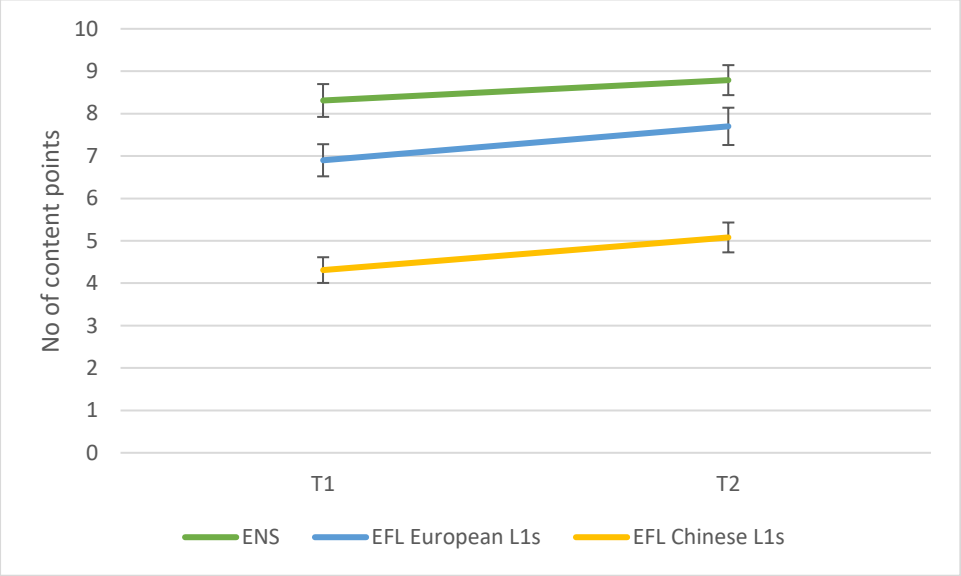

Figure 5. 13 Summarisation skills change across two time points

According to the mixed-design ANOVA, there was a significant main effect of time, *F*(1, 143) = 9.98, *p* = .002, *r* = 0.25. On average, the number of content points recalled across the whole sample increased from T1 (*N* =146, *M* = 6.51, *SD* = 2.98) to T2 (*N* = 146, *M* = 7.20, *SD* = 3.09). There was also a significant main effect of group, *F*(2, 143) = 36.815, *p* < .001, *r* = 0.45. A further post hoc test revealed that there was a significant difference between the ENS (*M* = 8.55) and the two EFL groups, the one with European L1s (*M* = 7.30), *p* =.025, and the one with Chinese L1s (*M* = 4.70), *p* < .001. A significant difference was also found between the two EFL groups (*p* < .001). The scores for summarisation skills yielded no significant interaction between time and group, *F*(2, 143) = .22, *p* = .800, *r* = 0.04. Despite the significant difference between the ENS and the EFL with European L1s as indicated in the post hoc test following the main effect of group in the mixed-design ANOVA, a closer investigation of results obtained on T2 data only showed that the significant difference between these two groups had disappeared by T2.

5.5.11 Spelling

The spelling error rate was based on the number of spelling errors made in the handwritten summaries. On average, the number of spelling errors increased from T1 to T2 in all three groups. The group of ENS obtained the smallest mean error rates from among all three groups, and the EFL with Chinese L1s made the greatest proportion of spelling errors at both time points (see Figure 5.14). The group of ENS deteriorated on average by 0.10 (*N* = 48, *SD* = 1.01, *Mdn* = 0, Range = -2.062–2.899); they recorded a mean rate of 0.90 (*SD* = .76) at T1 and 1.00 (*SD* = 91) at T2. This change, however, was not statistically significant, *t*(47) = -.67, *p* = .253, Cohen's *d* = -.097. The group of EFL with European L1s deteriorated by 0.14 on average (*N* = 49, *SD* = 1.33, *Mdn* = 0, Range = -2.86–4.292); their mean spelling error rate was 1.43 (*SD* = 1.39) at T1 and 1.58 (*SD* = 1.33) at T2. The difference between the two time points was not significant, *t*(48) = -.76, *p* = 226, Cohen's *d* = -.108. The group of EFL with Chinese L1s deteriorated by 1.19 on average (*N* = 48, *SD* = 2.35, *Mdn* = .71, Range = -2.44–8.86) as a result of an increase in error rate across T1 (*M* = 2.27, *SD* = 1.74) and T2 (*M* = 3.46, *SD* = 2.91), and change proved to be statistically significant with a medium effect size, *t*(47) = -3.51, *d* = -.507.

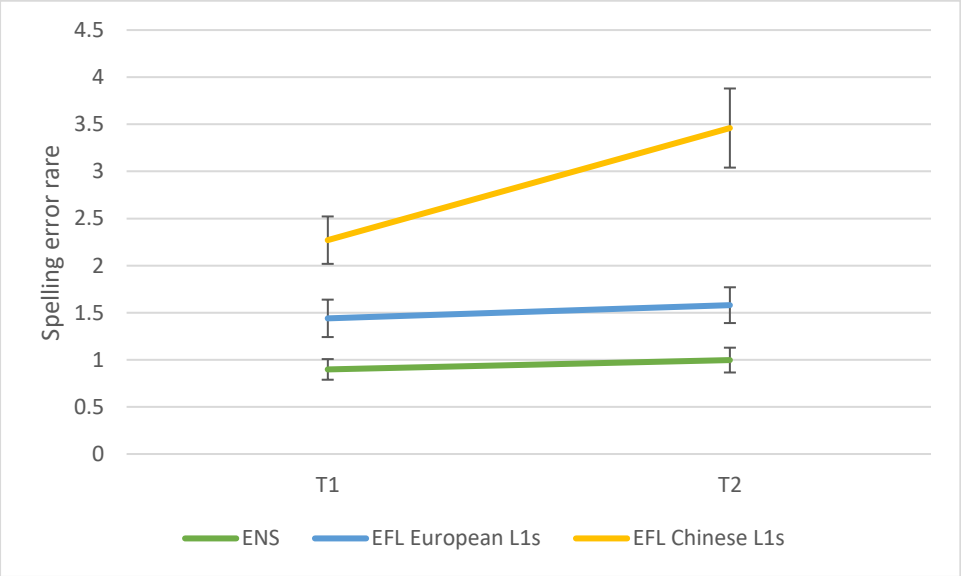

Figure 5. 14 Spelling error rate change across two time points

According to the mixed-design ANOVA, there was a significant main effect of time, *F*(1, 142) = 11.94, *p* = .001, *r* = 0.28. On average, the spelling error rate increased across the whole

sample from T1 (*N* = 145, *M* = 1.54, *SD* = 1.46) to T2 (*N* = 145, *M* = 2.01, *SD* = 2.17), which indicates a decrease in spelling proficiency across all three groups. There was a significant main effect of group, *F*(2, 142) = 22.611, *p* < .001, *r* = 0.37. A further post hoc test revealed a lack of significant difference between the ENS students (*M* = .98) and the group of EFL with European L1s (*M* = 1.51), *p* =.135; a significant difference was found between the groups of ENS and EFL with Chinese L1s (*M* = 2.87), *p* < .001, and a significant difference was found between the two EFL groups, *p* < .001. A significant interaction between time and group was found for this measure, *F*(2, 142) = 6.624, *p* = .002, *r* = 0.21. To further investigate these findings, a one-way ANOVA on the gains data was performed and a significant difference was found, *F*(2, 87.72) = 4.48, *p* = .014, *r* = .09 (see Figure 5.15). According to the Games-Howell post hoc procedure, there was a significant difference between the ENS (*M* = 0.10, *SD* =1.01) and the EFL with Chinese L1s (*M* =1.19, *SD* = 2.35), *p* =.012, and the difference between the EFL with European L1s (*M* = 0.14) and the EFL with Chinese L1s was also significant, *p* =.024.

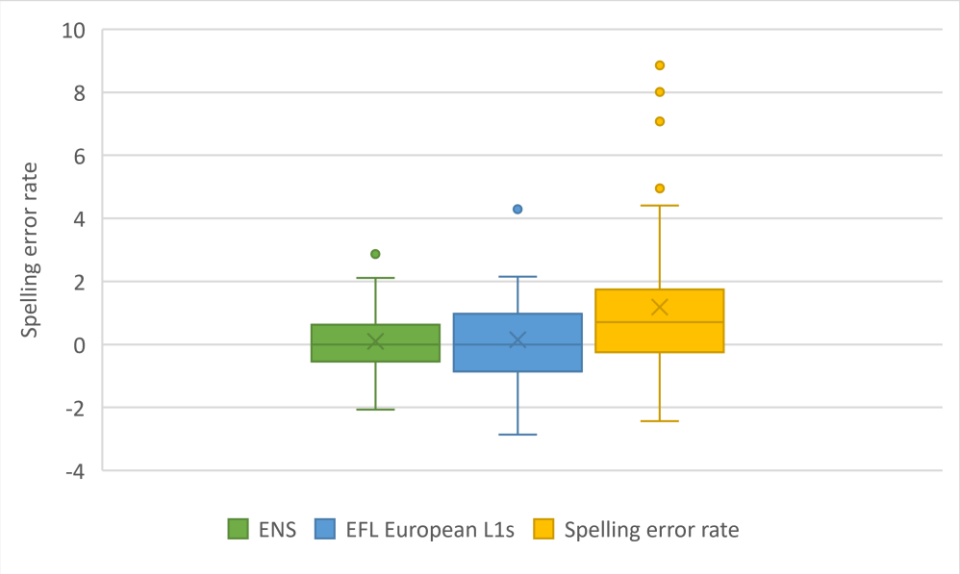

Figure 5. 15 Spelling error rate gains by language group

5.5.12 Elision

Finally, the results obtained in the two phonology tasks are presented. The first results came from the elision test, based on a scale between 0 and 20 with one point obtained for each correctly pronounced word following a relevant sound deletion. In general, all three groups performed better at T2 than at T1 in this task (see Figure 5.16), but none of them improved

to the point of reaching significance. The groups of ENS and EFL with European L1s performed very similarly at both time points and the EFL with Chinese L1s showed much lower performance when compared to the other two groups. The group of ENS gained on average 0.34 (*N* = 48, *SD* = 1.60, *Mdn* = 0, Range = -3–4). This group obtained a mean score of 17.56 (*SD* = 2.15) at T1 and 17.90 (*SD* = 1.93) at T2. The difference was not significant, *t*(47) = -1.44, *p* = .078, Cohen's *d* = -.208. The group of EFL with Chinese L1s improved on average by 0.40 (*N* = 48, *SD* = 2.18, *Mdn* = 0, Range = -4–6), as a result of the difference between 15.56 (*SD* = 2.76) obtained at T1 and15.96 (*SD* = 2.49) at T2. This change also was not statistically significant, *t*(47) = -1.26, *p* = .107, Cohen's *d* = -.182. The group of EFL with European L1s improved on average by 0.30 (*N* = 50, *SD* = 2.18, *Mdn* = 0, Range = -4–12) as a result of the difference between 17.62 (*SD* = 2.47) obtained at T1 and 17.92 (*SD* = 1.76) at T2, with a lack of significance, *t*(49) = -.97, *p* =.167,Cohen's *d* = -.138.

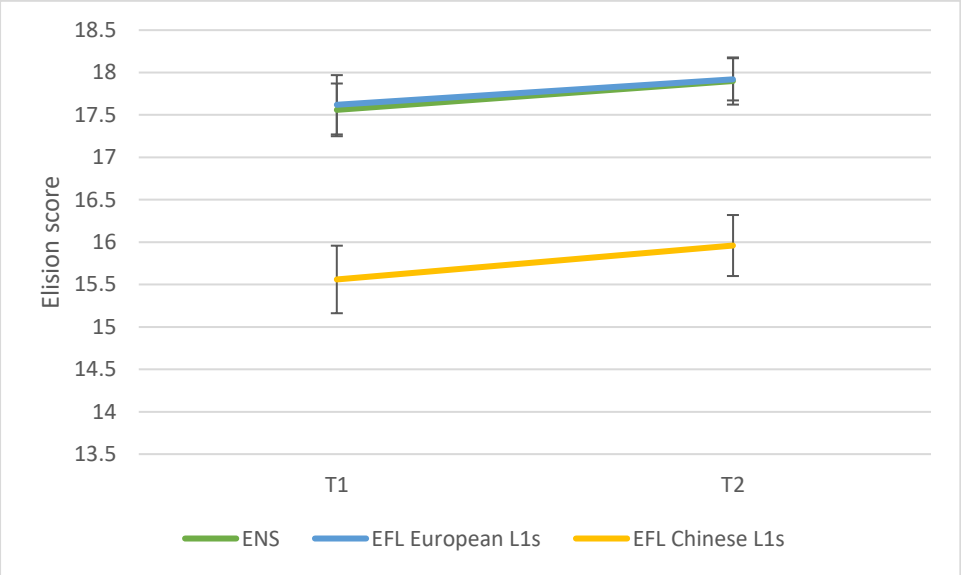

Figure 5. 16 Elision score change across two time points

In the elision task, a significant main effect of time was found, *F*(1, 143) = 4.259, *p* = .041, *r* = 0.17. On average, the mean score of the whole sample improved from T1 (*N* =146, *M* = 16.92, *SD* = 2.64) to T2 (*N* = 146, *M* = 17.27, *SD* = 2.26). There was a significant main effect of group, *F*(2, 143) = 15.138, *p* < .001. A further post hoc test revealed that there was no significant difference between the groups of ENS (*M* = 17.73) and EFL with European L1s (17.77), *p* = .994, but there was a significant difference between the group of ENS and the EFL with Chinese L1s (*M* = 15.76), *p* < .001, and a significant difference between the two EFL

groups of students, $p < .001$. No significant interaction between time and group was found, $F(2, 143) = .029$, $p = .972$, $r = 0.01$. In sum, all three groups improved but non-significantly, and the pattern of performance remained unchanged across the two time points: no significant difference existed between the ENS and the EFL with European L1s at any point, and the EFL with Chinese L1s showed significantly lower performance at both time points.

5.5.13 Rapid automatic naming of digits

Rapid automatic naming of digits was the second phonology task and is the final measure presented in this section. The analyses were performed on reading rates with the scores expressed as the number of digits read per second. On average, all three groups improved their reading rates across the two time points and the group of ENS obtained the greatest mean scores at both time points (see Figure 5.17). Two groups improved significantly across the two time points and these were the ENS and the EFL with European L1s. The group that improved its performance to the greatest extent was the ENS students, and they did so on average by 0.22 digits read per second ($N = 48$, $SD = 0.44$, $Mdn = 0.163$, Range = -0.702–1.215). Their mean scores increased from 2.79 ($SD = 0.52$) at T1 to 3.01 ($SD = 0.64$) at T2, with a medium effect size, $t(47) = -3.45$, $p < .001$, Cohen's $d = -.497$. The group of EFL with European L1s was the second best in terms of the amount of improvement and their average gain was 0.18 ($N = 50$, $SD = 0.38$, $Mdn = 0.227$, Range = -0.905–0.833). Their mean score at T1 was 2.56 ($SD = 0.48$) and at T2 it was 2.74 ($SD = 0.49$). This increase was highly significant with a medium effect size, $t(49) = -3.39$, $p < .001$, Cohen's $d = -.479$. Finally, the group of EFL with Chinese L1s gained on average 0.03 ($M = 47$, $SD = 0.74$, $Mdn = 0.128$, Range = -3.556–0.855), with their mean scores improving from 2.56 ($SD = 0.75$) at T1 to 2.59 ($SD = 0.44$) at T2; this, however, was not significant as indicated by the *t*-test, $t(46) = -.27$, $p = .393$, Cohen's $d = -.040$.

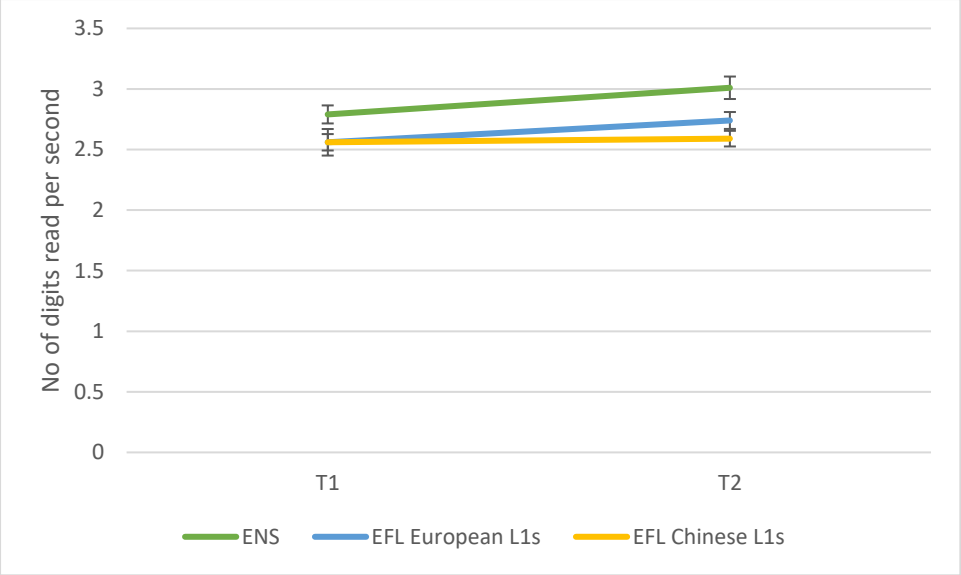

Figure 5. 17 Rapid naming rate change across two time points

According to the mixed-design ANOVA, there was a significant main effect of time, $F(1, 142) = 10.23$, $p = .002$, $r = 0.26$. On average, the reading rate increased from T1 ($M = 2.64$, $SD = .60$) to T2 ($M = 2.78$, $SD = .56$), which indicates an overall improvement. There was a significant main effect of group, $F(2, 142) = 5.55$, $p = .005$, $r = 0.19$. A further post hoc test revealed that there was a significant difference between the ENS ($M = 2.90$) and the EFL with European L1s ($M = 2.65$), $p = .046$; a significant difference was found between the group of ENS and the group of EFL with Chinese L1s ($M = 2.57$), $p = .008$, and no significant difference was found between the two EFL groups of students ($p = .675$). The time by group interaction proved to be non-significant, $F(2, 142) = 1.65$, $p = .196$, r = 0.11. Despite the fact that the post hoc test following the main effect of group showed a significant difference between the ENS and the EFL with European L1s, these two groups were not significantly different in their performance at T2 as indicated by another post hoc test performed on T2 data only.

5.6 Summary of T2 findings

In general, all three groups improved on nearly all measures over the period of one year. A decrease in performance was observed in all three groups across two measures: Nelson-Denny untimed reading comprehension and spelling error rate. In two other measures, reading rate and total number of words produced in summaries, the groups showed a mixed

pattern of improvement and decline. What is interesting, in these two measures the same pattern of performance is observed, as the ENS and the EFL with Chinese L1s recorded drops in their scores while the EFL with European L1s were the only ones who improved. In other words, the EFL with European L1s recorded decrease in performance in two measures only (untimed reading comprehension and spelling error rate) whereas the groups of ENS and EFL with Chinese L1s recorded decrease in performance in 4 measures (untimed reading comprehension, spelling error, reading rate, and total words in summary).

The group of EFL with European L1s improved to the greatest extent. This is indicated by the number of tasks where they obtained the greatest and significant effect size. This is true for five measures: vocabulary, vocabulary response time, grammar, Nelson-Denny untimed reading comprehension, and writing rate. The group of EFL with Chinese L1s recorded the greatest and significant effect size for two measures only: reading fluency and summarisation skills. They improved significantly in two further tasks, vocabulary and vocabulary response time, both yielding a medium effect size. The group of ENS improved significantly in two measures of vocabulary, reading fluency, and the speed of processing in rapid naming of digits.

All the main effects of group in the mixed-design ANOVAs reached significance. According to the post hoc tests on cumulative T1 and T2 scores in all three groups, the ENS students performed significantly better than the EFL with Chinese L1s in *all* measures. The same is true for comparisons between the two EFL groups: there were significant differences in *all but one* of the measures, with the exception being rapid naming of digits. The pairwise comparisons between the ENS and the EFL with European L1s following the main effect of group in the mixed-design ANOVAs showed that there was no significant difference between the two groups in the following seven measures: 1) Which English grammar test, 2) sight word efficiency, 3) Nelson-Denny untimed reading comprehension, 4) writing rate, 5) total number of words in summary, 6) spelling error rate, and 7) elision. The main effect of time was significant in all but the following measures: reading rate, Nelson-Denny untimed reading comprehension, and number of words in a summary.

The time by group interaction reached significance for only one measure, the spelling error rate. The direction of change in this task, however, indicated a deterioration in relation to T1 findings. All three groups demonstrated higher error rates at T2 than at T1. The significant

time by group interaction therefore indicated that the group of EFL with Chinese L1s experienced a greater drop in their performance than the other two groups of students.

Chapter 6: Discussion

This longitudinal study was an investigation into knowledge of English and literacy skills in a heterogeneous population of university students at a UK university. Its primary goal was to better understand the differences between British home students who pursue their courses in their first language, and international students, who study with English as a foreign language. The second aim of this study was to extend the findings from Trenkic & Warmington (2019), and to find out if large and significant differences found previously between British home students and international students from China generalise to other student populations who are very different in many aspects. To achieve this goal, Chinese L1 students were compared against another group of international students, who spoke European first languages and were very different from Chinese L1 students, not only linguistically but also culturally and geographically. In addition to the main goals stated at the start of the study, the results make it possible to shed some light on the relevance of the level of proficiency in English set by universities in a diverse and multilingual university student population. Two main research questions were stated in this investigation, the first one about the differences in language and literacy skills at the start of university between British home and international students, and between two diverse groups of EFL students. The second research question pertained to language development over the span of one academic year.

To answer these questions, a task battery was developed and administered to three groups of undergraduate students at the start of university: native speakers of English (*N* = 59), EFL students with European L1s (*N* = 60), and EFL students with Chinese L1s (*N* = 59). The battery consisted of tasks assessing cognitive abilities (intelligence and working memory), linguistic knowledge (knowledge of vocabulary and grammar), literacy skills (skills in reading and writing), and phonological skills. The data obtained at this time point was analysed with ANOVAs and planned contrasts to answer the first research question about between group differences in all measures at the start of year one. The same task battery (excluding cognitive ability tasks) was administered after one year to a subsample of participants who returned for testing at the beginning of their second year. This group consisted of ENS (*N* = 49), EFL speaking European L1s (*N* = 50), and EFL speaking Chinese L1s

(*N* = 48). The data obtained at both time points were analysed with *t*-tests, and mixed-design ANOVAs to see how much language has changed in all three groups, which groups improved and to what extent, and whether the differences observed at the first time point had disappeared over the course of one academic year.

This chapter will discuss the findings in relation to both research question (6.1 and 6.2), provide potential implications of these findings on academic achievement (6.3), and propose some practical and pedagogical solutions to some of the issued identified (6.4). It will finally discuss limitations and future research avenues (6.5).

6.1 Research question 1

This section discusses the main findings that are based on the data collected at T1, at the start of year one at university:

RQ1. How much do knowledge of the English language and literacy skills differ at the start of the first year at university between:

a) Home students who speak English as their first language and international students who speak English as a foreign language (EFL)?
b) Those speaking European L1s and Chinese L1s?

The results showed that, despite no differences in fluid non-verbal intelligence and working memory, large and significant differences existed between undergraduate native speakers of English and the whole EFL group at the start of the first year at a UK university. Large and significant differences were found across *all* measures obtained in language, literacy, and phonological tasks. These results corroborated previous findings on the initial differences between home and international university students in vocabulary (Elder et al., 2007; Morris & Cobb, 2004; Trenkic & Warmington, 2019), and reading and writing (Devos, 2019; Harrington & Roche, 2014; Trenkic & Warmington, 2019). Large and significant differences were also found in the measures of phonological skills, speed of language processing, and spelling, that is, skills that underpin language and literacy development, which has received very little attention in research so far. In addition, large and significant differences were found in grammatical knowledge, which is an aspect of language that has not previously been investigated with regards to language and literacy skills of international students.

These findings suggest that international students, on the whole, are a group that linguistically differ greatly from native speakers of English at the start at university.

The findings also confirmed large and significant differences between the group of British home and Chinese L1 students at the beginning of the first year at university (Trenkic & Warmington, 2019). In the present study, these two groups differed in *all* measures of language, literacy, and phonological skills, despite no significant differences in their cognitive abilities. What is more, students speaking Chinese L1s obtained the highest mean scores in the measures of non-verbal fluid intelligence and working memory across all three groups, demonstrating that their cognitive abilities were not likely to influence the results obtained. The differences between the British home students and the EFL speakers with Chinese L1s corroborated the magnitude of difference found in Trenkic & Warmington (2019).

The Chinese L1 participants in this study knew significantly less words than the group of ENS and the maximum score obtained by Chinese L1 participants in the vocabulary task was below the average score in the ENS group. The same is true for grammar as none of the Chinese L1 students were able to obtain a maximum score in the grammar test and the variability in their scores was the greatest from among all three language groups. Chinese L1 students' timed reading comprehension score was half of those obtained by the ENS group, and the same was true about their summarisation skills, as their score here was half of those obtained by the ENS. The Chinese L1s were also significantly slower in all language processing measures, had significantly weaker phonological skills, and made significantly more spelling error when compared to the group of ENS. The high level of variability in scores from the Chinese L1 sample may suggest that this group itself was very heterogeneous even though they shared the same linguistic and cultural background. The reason behind this behaviour is beyond the scope of this work but very interesting for future investigations.

The first major contribution to the field made by the present study is the finding that shows large and significant differences between the international students speaking European L1s and those speaking Chinese L1s. Some differences between these two groups have been observed in previous studies on faculty perceptions of students' abilities in English (Mayor, 2006; Senyshyn et al., 2000; Trice, 2003). However, the present study fills the theoretical

and methodological gaps by selecting participants according to students' first languages and comparing actual skills in English in a comprehensive way with a battery of assessments. According to the results obtained in the ANOVAs and planned comparisons, large and significant differences were found between the two EFL groups of students in all but one task. The only task where no significant difference was found was the rapid naming of digits. This result may stem from the fact that digits are among the most learned, practiced, and consequently automated lexical items, which allowed Chinese L1 students to perform as quickly as other EFL students.

Another original finding that is based on data obtained in this study was that the difference in language and literacy skills between the group of ENS and EFL with European L1s at the start of university was not as great as the difference between the ENS and the group of EFL speaking Chinese L1s. This means that the large and significant differences observed between ENS and students speaking Chinese L1s do not necessarily generalise to all groups of international students. More specifically, these findings cannot be generalised to EFL students whose first languages belong to the same family as English, and who are culturally and geographically closer to English. Even though the group of European L1 speakers obtained lower mean scores in almost all measures, these differences were not significantly different from ENS students in almost half of the measures.

No significant differences were found between the ENS and EFL speaking European L1 in reading fluency, untimed reading comprehension, number of words produced in the writing task, spelling, and two tasks assessing phonological skills. This shows that the large and significant differences between the home and the whole group of international students were driven by the Chinese L1 students. The most important implication of this finding is that international students cannot be treated as homogeneous group, neither in the university studies context, nor in research. Future studies on language and literacy skills need to account for participant linguistic and cultural background as the results can be obscured in the way observed above: large and significant differences existed between the ENS and international students, but some patterns emerge when students are selected according to certain characteristics.

6.2. Research question 2

The study also tracked language development over the course of one year, in native speakers of English and different groups of international students. The aim was to see if lack of significant improvement, large between-group differences after one year, and lack of closure of the gap in relation to native speakers of English observed in a sample of Chinese L1 students (Trenkic & Warmington, 2019), can also generalise to students speaking European L1s. The main goal of this investigation was to understand if the level of English international students began their courses with allowed them to catch up with the native speakers of English in a timely manner. The closure of gap by the second year at university would suggest that, despite some initial differences, the level of English international students begin university with provides equal opportunities and a level playing field for those studying in their second language. The following research question was stated:

RQ2. How much do knowledge of English and literacy skills change in the first year at university in home students and in international students speaking Chinese and European L1s? Do international students close the gap on any measures?

The results obtained across the two time points showed positive trends in language change in almost all measures investigated. Contrary to the initial prediction, the mean scores in two measures: reading rate and the total number of words produced in the writing task, did not improve in all of the groups by the beginning of the second year at university. The decrease in reading rate across the two time points was observed in two groups: ENS and EFL with Chinese L1s, but the EFL with European L1s recorded a small and non-significant improvement. The performance in this task could have been impacted by different modes of task administration across the two time points. In T1 data collection, the participants read the text from paper in a lab, whereas at T2, the same text was read on the screen in a Zoom meeting. It is very likely that this affected the expected improvement in a negative way because research shows that reading from screen leads to eye fatigue and consequently slows down the speed of reading (Dillon, 1992; Nielsen, 1997, 2010). The fact that the EFL group with European L1s improved despite the difficulty caused by reading from screen may suggest that the magnitude of their improvement must have been much greater than the

non-significant improvement recorded at T2. In other words, if they were asked to read the text on paper at T2, they would be likely to improve to a greater extent. Students speaking European L1s not only overcame the challenge of a different mode of text presentation but also recorded a small improvement.

Another measure where the mean scores did not improve across the two time points, contrary to the initial assumption, was the number of words produced in the writing task. A drop in the average number of words produced was observed in two groups, ENS and EFL with L1 Chinese L1s, but the group with European L1s recorded an improvement. Here, this unexpected pattern of performance can be explained by different conditions in which the task was performed and the amount of control over participants while performing it on Zoom. In T1 data collection, all the participants were provided with the same conditions with identical seating arrangements and stationery provided for them. However, in T2, participants were asked to prepare their own stationery before the meeting began and wrote in their home environment, which could have introduced a more relaxed attitude. If this interpretation was correct, we would expect all three groups to display similar behaviour, which was not the case here as the group with European L1s did, in fact, improve their mean score at T2.

Another explanation for the results obtained in this measure may, again, originate in the adverse effect of reading from screen. The written task was based on reading a text off a screen, which was previously presented on paper. Research shows that reading from screen can not only affect the speed of reading, but also text comprehension. For example, a recent meta-analysis by Kong et al. (2018) showed that reading on paper was better than reading from a screen in terms of text comprehension. Lower comprehension rates could lead to remembering less and, consequently, a lower number of words produced. The group of EFL students with European L1s was the only one that overcame the limitations of reading from screen, as they improved their reading speed and by extension, they could have increased comprehension and consequently included more words in their summaries. The improvement in the number of words produced by the European L1 sample was positive and not significant, but the level of their improvement, again, may be underestimated here. The connection between the two measures: reading rate and number of words, and the fact

that only one group improved in both measures makes the adverse effect of reading from screen a more plausible explanation here.

The final measure where a negative change in performance across the two time points was recorded was spelling, with all three groups making more spelling errors in their summaries at T2 when compared to T1. However, the results for this measure were difficult to predict. It was deemed possible that an initial increase in the number of spelling errors could occur as a result of new vocabulary acquisition. Consistent with this prediction, it would be expected that the spelling would deteriorate the most in those groups that acquired the most vocabulary. This, however, was not supported by data obtained in the vocabulary task, as the group that recorded the most significant increase in spelling error rate, was in fact the group that acquired vocabulary at lowest rate, which was the Chinese L1 students. The other two groups, on the other hand, showed a non-significant increase in the number of spelling errors, and their vocabulary gains were the highest across all three groups. Therefore, the prediction that spelling rate would go hand in hand with vocabulary acquisition, does not hold here.

Another explanation may originate in differences in writing between the alphabetic writing system used in English and the logographic writing system used in Chinese. It was explained in the literature review that words such as *conversation* and *conservation* can pose difficulties for Chinese L1 students because they are very difficult to discriminate due to the number of letters and syllables, and consequently, it is easier to misspell them. This claim finds some support in the data obtained in the handwritten summaries at T2. Chinese L1 students made many spelling errors in long, and multi-syllable words including the following misspelt examples: *chocolate, luxurious, empire, continent, remarkably, monarch, extremely, appreciated, convenience, wealthiest, monopolised, beverage, ingredients, widespread*. Some of those words are of relatively high frequency, such as *chocolate*, and expected to be known by Chinese L1 students upon arrival. Therefore, it is not likely that spelling occurred in a newly acquired vocabulary item. Still, the analysis of misspelt vocabulary items in summaries has not been carried out in a systematic way and this explanation needs to be treated with caution.

In the remaining measures, all three groups recorded an improvement across the two time points. The group of EFL students with European L1s made the greatest

improvement among all three language groups, in terms of both the magnitude of improvement and the number of measures on which they improved significantly. The EFL students with European L1s recorded the largest, compared to all three groups (as indicated by Cohen *d's*), and most significant improvement in the following measures: vocabulary, vocabulary response time, grammar, timed reading comprehension, and writing rate. If we add the two measures where the EFL students with European L1s was the only one that improved (reading rate and number of words), and assume that their improvement there was underestimated, they are the group that improved the most significantly among all three groups on seven measures. This constitutes half of the measures administered in the task battery. This improvement allowed the group speaking European L1s to close the gap in relation to the group of ENS on *almost all* remaining measures that existed at the beginning of university.

This was true for grammar, where a significant improvement allowed European L1 students to close the marginally significant difference that existed between them and the ENS at the beginning at university. Another gap that closed in relation to the ENS was in the measure of timed reading comprehension, where European L1 students improved significantly and the most compared to all three groups, making the closure of the highly significant gap that existed between them and the ENS at the start of university possible. Next, the largest and highly significant improvement in writing rate also led them to close the gap with the ENS by the beginning of the second year at university. The gap that existed in summarisation skills also closed thanks to a relatively small but significant improvement. Finally, the gap in reading rate also disappeared across the two time points, but the effect here was mainly because of the drop in the performance in the group of ENS. By the beginning of the second year at university, the group of EFL students with European L1s was not distinguishable from the ENS in all but two measures: vocabulary and response time in the vocabulary task. These two gaps did not close despite a significant and large improvement recorded by the group of European L1 speakers.

The group of EFL students with Chinese L1s improved significantly and more than the other groups in only two measures across the two time points: reading fluency and summarisation. They also improved significantly in vocabulary and vocabulary response time. Despite relatively large and significant improvement, they were not able to close the

gap in these measures in relation to the group of EFL students with European L1s. By the beginning of the second year at university, the group of students speaking Chinese L1s were still significantly different across all measures when compared to the group of ENS, and significant gaps existed between both EFL groups in all but one measure (rapid naming of digits).

At the same time, the group of ENS were also improving their knowledge and skills making the task of catching up even more challenging for English learners. This group recorded the greatest and most significant improvement among all three groups in rapid naming of digits, and also improved significantly in vocabulary task, response time in vocabulary task, and reading fluency.

Participants' performance in one of the tasks is standing out and this was the score and the reaction time improvement in the vocabulary size task. All three groups improved significantly and considerably with large effect sizes as indicated by Cohen's *d,* in the order of magnitude: the European L1s students (*d* = -.815), EFL with Chinese L1s (*d* = -.487), and the ENS (*d* = -.399), as for the score obtained in the VST which was based on the number of correct answers out 140. The same is true about the speed for providing the correct answer in this task: the group of EFL with European L1s improved to a greatest extent (*d* = .620), next in the order of magnitude was the group of ENS (*d* = .597), and finally the group of EFL with Chinese L1s (*d* = .493). Can these large effect sizes be explained by the effect of familiarity with the task, as the same version was administered at both time points? Is it possible that students memorized some of the unknown vocabulary items presented in the first test administration, checked their meaning after the test out of curiosity, and remembered them by the T2 testing session which led to an increase in test score at T2? Or, has it improved because of more lenient testing conditions at T2 where students were taking the test in home environment?

It is highly unlikely that students were cheating at T2 as they had their mobile phones switched off and their computer screens were shared with the researcher, which limited their access to digital or online dictionaries. What is more, searching for the meaning of unknown words elsewhere would increase the reaction times for providing the correct answers, and the opposite trend was found with their reaction times being in fact much

faster when compared to T1 performance. When it comes to searching for and memorizing the meaning of unknown vocabulary items following the first exposure to these words at T1, if this was the case, the group of EFL with European L1s would demonstrate the strongest effect of this behaviour with a very high effect size for the gain in the number of correct answers. Ironically, this would further confirm this group's strength not only in new vocabulary learning but also retention. It can be therefore concluded that the VST, despite its limitations, was still suitable for the purposes of the present study context and the group with European L1s seem to be the strongest one from among all three groups of participants in their vocabulary development.

In sum, these findings show that the group of EFL students with European L1s began university with a level of English that allowed them to catch up with the native speakers of English on almost all measures of linguistic knowledge and skills. The tasks that proved to be the most challenging were the ones examining vocabulary knowledge and response times in the vocabulary task. Despite very large and significant improvement in vocabulary score (t = -5.70, Cohen's d = -.815, p < .001), and response time in vocabulary task (t = 4.34, Cohen's d = .620, p < .001), the group of EFL students with European L1s did not manage to catch up with the group of ENS. It is worth looking closer at the results of the vocabulary task here as they show that the sample was typical when compared against findings in other studies. Also, the vocabulary size of 8,000–9,000 word families suggested for pursuing university studies does not seem to be sufficient because it does not allow further development of vocabulary at a rate typical for English learners in second language context.

According to the estimated vocabulary size based on the VST algorithm, the group of ENS improved at a rate of around 700 word families over one year (from 16,415 word families at T1 to 17,085 at T2) which is slightly lower than the expected yearly growth of 1,000 word families in this type of population (Goulden et al., 1990). The group of EFL students with European L1s increased their vocabulary at rates that were above the expected yearly growth of 1,000 word families in English learners in a second language learning environment (from 11,393 word families at T1 to 12,784 at T2) (Milton & Meara, 1995). The EFL with Chinese L1s, on the other hand, acquired on average 400 word families which is well below

what is expected of English learners in an immersion context (from 7,441 word families at T1 to 7,857 at T2).

Even though the Chinese L1 students were slightly below the minimum vocabulary range of 8-9,000 word families suggested for learning at English medium universities, it did not facilitate suitable vocabulary development during the first year at university. This was because their improvement was well below the growth rates expected of English learners in a second language context. An initial vocabulary size of 11,000 word families demonstrated by the group with European L1s seemed to be a more suitable starting point at an English medium university, as it facilitated growth that was beyond of what is expected of second language learners. The vocabulary size of 11,000 word families at the beginning of university seemed to be a springboard that allowed international students growth at a rate comparable to those in native speaker of English. Even though the group with European L1s did not catch up with the native speakers of English in terms of vocabulary knowledge by the beginning of the second year, their vocabulary growth was higher than a typical growth rate expected from international students in the second language learning environment. This puts them on the right course it terms of vocabulary development and ultimate closure of this gap in relation to native speakers of English.

Despite some initial concerns over Zoom, this tool proved to be a reliable research tool that was able to mimic an in-person interaction. This claim found support in the patterns of performance in all three groups across both time points. Apart from the two measures mentioned above (reading rate and the number of words in the writing task), where the results were likely affected by the adverse effect of reading from screen, no other measure seemed to be impacted in a negative way. There were no disruptions or sound delays due to technology or the environment, which confirms Zoom's reliability stressed elsewhere in the literature (Matthews et al., 2018). Also, the tasks that were time and sound sensitive were not affected by Zoom in any adverse ways. This conclusion is based upon looking at the patterns of performance in these tasks across the two time points. The mean scores in these tasks improved across both time points, as expected, and the patterns of between group differences that existed at T1 were reflected in the findings obtained at T2. Had Zoom affected the results in a negative way, these results would not have been so consistent. It can therefore be concluded that Zoom was able to mimic in-person interactions and it

seems to be suitable for behavioural and perceptual studies, similar to the present one, as well as interview-based studies.

Large and significant differences between the two international groups of students revealed by this study may originate in the linguistic distance from English. This is because students speaking Chinese L1s have to overcome some learning burden that is not present in students speaking European L1s. This includes the differences in writing system (Ross & Ma, 2009), conceptualising vocabulary (Ma & Kelly, 2009), and differences in phonology and grammar (Li & Thompson, 2003). However, there are several other factors that may be responsible for the findings. The results obtained in this study may be impacted by the cultural and educational differences between students from China and Europe. Students from China are not usually familiar with the expectations for written assessment, criticality, and expressing opinion. This group was also found to have problems with acknowledging sources and referencing in writing (Jin & Cortazzi, 2006). These problems may originate in Chinese educational system and creates a steeper learning curve when compared to those from Europe. Also, the geographical distance from the UK gives more opportunities to those whose place of residence is Europe, to travel more extensively and learn more about the host country culture and its educational system. Other factors may include the amount of exposure to English prior to arrival, but also after enrolment as the data obtained in this study showed that Chinese L1 students seem to use their first language to a greater extent than students speaking European L1s on daily basis.

6.3 Implications for academic achievement

Literature from the context of children with limited proficiency in English shows that English learners can compete with native speakers academically only if they catch up with them linguistically. No corresponding research exists in the HE context. However, if the same was true for HE students, undergraduate students in the UK would need to arrive with language that allows them to catch up with the native speakers of English over the first year, before they start earning grades that count towards their final degree classifications. Consequently, English medium universities would need to set the level of proficiency in English at that allows international students rapid language improvement and closure of the initial gap between them and native speakers of English. Setting the minimum language level following

this rationale would be perceived as the right and ethical approach as international students would be able to unleash their full potential while at university.

In the recruitment process, in most of the cases universities use high stake English language proficiency test such as IELTS. The score obtained in this test is placed on a 9-point scale and according to IELTS test developers, the IELTS cut score of 7.0 is suitable for linguistically less demanding, and 7.5 for linguistically more demanding courses. However, universities do not always follow this guidance and set lower cut scores as the decision may depend on other qualifications a student arrives with, or to secure the right number of enrolments. In case of participants taking part in this study, IELTS was demonstrated as entry qualification in 22 (out of 60) international students speaking European L1s and almost all students with Chinese L1. A closer examination of the level of IELTS in both samples shows that the range of IELTS scores in the group speaking European L1s was 6.5–8.5 with a mean of 7.5. In the sample of students with Chinese L1 the range in IELTS was 5.0–8.0 with a mean of 7.0. Several students in this sample were admitted with IELTS as low as 5.0, and bypassed the minimum cut score by attending pre-sessional courses. This shows that in case of students with Chinese L1s, they have not reached the IELTS of 7.5 suggested by test developer for students on linguistically demanding courses. Also, the proficiency in English they arrived with did not allow them to close the initial gap with British home students and other international students. Those speaking European L1s, on the other hand, have reached the cut-off score suggested by the test developer, and this group managed to close the initial gaps that existed between them and British home students.

The fact that students speaking European L1s reached the cut-off score of IELTS suggested by the test developer and managed to close the gap with the British home students may suggest that the IELTS cut score of 7.5 is adequate to begin university with as it may allow international students to close the initial gap with native speakers of English. This conclusion, however, needs to be treated with caution as IELTS data comes from only 30% of the sample with European L1s. At the same time, it may look as if the IELTS cut score set by universities may not be aligned well with the linguistic demands of courses. When setting the cut scores below those set by the test developer, universities would need to be prepared to mitigate the adverse effects it can have on language skills with an effective language provision for international students.

Given the importance of proficiency in English with which international students begin university with, and that the chances of obtaining a good degree classification increase with an increase in knowledge of English, it is likely that the group of EFL with European L1s have greater chances to obtain a better degree than the group of EFL with Chinese L1s. Knowledge of vocabulary, grammar, and skills in reading and writing are central in ultimate academic achievement. Since Chinese L1 have these skills significantly lower when compared to EFL with European L1s and native speakers of English, they may be responsible for their underachievement in relation to native speakers of English and other international students reported in the literature (Morrison et al., 2005).

6.4 Pedagogical and practical implications

According to the findings, students with Chinese L1s arrive at English medium university with less-well developed knowledge and skills in English when compared to students from Europe. These findings may be therefore important for university language provision centres when targeting students that may need language provision and working on relevant policies. Since the findings show that European L1 students are on the right course when it comes to learning English and the language in Chinese L1 students is still developing, the latter may be prioritised when it comes to language support. The main issues highlighted in research on linguistic adjustments to university in Chinese L1 students were understanding technical vocabulary, reading, and writing (Evans & Morrison, 2011; O'Connell & Resuli, 2020; Yen & Kuzma, 2009), and speed of language processing (Zhao & Mawhinney, 2015). These findings seem to be consistent with the findings obtained in this study as these aspects are less-well developed in Chinese L1 students when compared to other international students. Therefore, these aspects should be prioritised when considering language provision for international students.

An effective language provision may not be the only solution in helping to improve English in Chinese L1 students. There are other strategies that can be implemented. Since it was found that Chinese L1 students socialise mostly through their first language, more opportunities can be created for them in order to mix with home, but also other international students. The university where this study was conducted seems to be very proactive in this area through organising social activities and schemes aimed at international

students such as buddying scheme. Under this provision, a newly arrived international student is paired with a more experienced fellow students who helps in the process of accommodation to a new environment. Students themselves should be encouraged to continue learning English while at English medium university. They can do so by mixing with other English users, socialise using English, but also by exposing themselves to English more through extensive reading of English books, reading newspapers, watching English movies, and engaging in other aspects of the British life and culture.

Still more needs to be done and given the crucial role of vocabulary and speed of language processing, students from China may be allowed to spend more time in examinations and to use dictionaries; these ideas have been proposed by other researchers too (Mendelson, 2002). Universities across the UK have slowly started to accommodate for international students' needs by implementing these solutions. For example, international students at the university of Birmingham are allowed to bring dictionaries to their exams. In the exam period, they are advised by email that, if their first language is not English, they can bring a dictionary with them. The exact conditions of using dictionaries are presented in a set of rules. For example, an Italian speaker may use an Italian-English dictionary in a French exam but is not allowed to use a dictionary in an Italian exam. Students are also asked to fill out a Dictionary Approval letter, bring it to the exam and hand in to the exam invigilator. This seems to be a reasonable solution considering the finding according to which vocabulary is the most challenging to develop, irrespective of the first language. It is also available for those who feel that they need it and communicated to students in a friendly and polite manner. Despite these advantages, it still may bring some administrative burden for universities. The program administrators must ensure that the message was put across to all the students, the exam papers need to reflect the use of a dictionary in form of additional column to tick, the exam invigilators are responsible for additional checks of the dictionary approval letters in their procedure, and they must inspect the physical copies of dictionaries brought by students to prevent unethical behaviour. Despite its many advantages, allowing dictionaries in exams brings challenges to universities and the exact cost benefit analysis would be ideally subject of a future investigation.

Universities are expected to make more informed decisions while setting language requirements for international students as this study suggests that some international

students, despite meeting these requirements, do not reach the expected standards needed to compete with native speakers of English and other international students. It was pointed out that admission offices' decisions are not always justified by relevant evidence and that universities in many cases compete to secure the target number of international enrolments by admitting students whose language is below the benchmarks stipulated by the standardized test developers. This practices not only undermine international students potential and chances, but also can have some long-term consequences such us declining value of internationally obtained degree and the quality of EMI education in general. The increase of internationalization has many positive results and benefits the whole economy, many institutions, and individuals. However, it is important to be vigilant about any possible negative consequence and this provision should be carefully scrutinized to bring the intended benefits.

Universities can, for example, engage more in research on the relationship between the level of English and students' academic attainment. This can be done by collecting data on the use of English as student records do not provide such information. English-medium universities can also monitor the level of academic attainment in international students and compare it to those in home students.

6.5 Limitations and future studies

This study was conducted in atypical conditions because of lockdowns that affected the second half of the first year at university. Even though it was found that both samples of international students spent comparable amount of time abroad, less is known about the way in which the circumstances impacted learning and language development. The conditions of lockdowns, social distancing, distant learning, are novel occurrences that had not happened before and it is not known what impact it had on each of the three groups of students, and whether it affected them to the same degree. This study therefore should be replicated in more standard conditions for further validation of the findings.

One of the limitations of the present study was that the two EFL groups were different in terms of proficiency assessed with the standardised proficiency tests they arrived with. The group of EFL with European L1s had significantly higher score than the group of EFL with Chinese L1s. Even though the Chinese L1s students reached the level of proficiency

stipulated by the university, it is not known what the results would look like if the two groups did not differ in terms of qualifications they arrive with. Future studies could ideally match participants according to the level of results in standard proficiency tests such as IELTS. Also, the limitation in controlling for participants' background needs to be addressed in a future study. Participants in each language sample would ideally comprise a more uniform group not only in terms of their first language, but also nationality status as this would allow for better control over the confounding variables such as the pedagogical and economic factors that impact the instruction and learning English. The future comparisons are advised to be based on samples of students from the same European or Asian country.

Upon probing into the exposure to English, it was found that EFL students speaking Chinese L1s were exposed to English to lesser extent than the other EFL group as they reported having more friends speaking their own first language. Therefore, language development investigated in the present study could have been affected by number of friends having the same L1. Future studies should take the language use patterns into consideration, but also many other that could be influential such as, ways of learning English, learning strategies, access to English-media prior to arrival and the challenges posed by artificial intelligence. This final aspect, namely the use of artificial intelligence, can involve instant translations of extended texts including, for example, a whole scientific article that needs to be read for a seminar. Some anecdotal evidence from my personal experience while teaching students from China suggests that the use of translation software is a very common practice among this cohort. However, no study so far investigated the impact of artificial intelligence on learning and language development making it an interesting future research avenue. Even though the present study used a rigorously selected sample, there are still of other factors that could have impacted the results.

Finally, to understand the relationship between language and academic attainment, this study could benefit from additional analysis, namely, examining the relationship between the initial language skills and academic attainment after one year. This was, indeed, one of the research questions and the study was designed to obtain students' grades. For this reason, they were asked their permission to obtain their marks. Despite our request made within the relevant university services, these data were not shared with us. Even if it was, the data would not be complete and reliable. This is because the data we requested

concerned participants grades in the first year at university. Because of Covid 19, most of the exams and assessments were cancelled and students were engaged in alternative assessments. Even if these data were available for analysis, it would not reflect students' true abilities. What is more, future studies need to compare not only between group differences, but also look comparisons of individual students' scores.

Research into English and literacy skills of university students is needed now more than ever. The new policies implemented in the UK as a consequence of Brexit forced European students to apply for students' visas, increased the tuition fees, and caused additional difficulties for European students while searching for graduate jobs. These developments led to a sharp decline in the number of European enrolments: from over 66 thousand in the academic year 2020–2021, to just over 30 thousand in the following academic year (HESA, 2023). This 50% drop in the number of European students have several negative implications on language and literacy skills in the UK and elsewhere. Firstly, the places at the UK universities reserved so far for European students need to be filled by other international students. According to the latest statistics, the loss of European cohort was compensated by increasing the numbers of non-Chinese Asian students whose numbers raised from over 43 thousand in the academic year 2020–2021 to almost 67 thousand in the following academic year (HESA, 2023). These international students speak first languages that are very different from English, which introduces even more variation in the range of language and literacy in the receiving institutions.

The second implication of the new post-Brexit reality is further expansion of English medium provision in the European continent. The universities in the EU have quickly adapted to cater for the needs of those European students who wish to pursue their Education with English as a medium of instruction. For example, Germany's foreign enrolment reached record high with an increase of 8% in 2021–2022 academic year. This was the second consecutive year of growth after the 1.5% year-over-year gain in 2020–2021 (ICEF Monitor, 2022). It is important to study the development of English in non-English speaking countries where students have limited exposure to the target language and less opportunities to learn when compared to their peers in the traditionally English-speaking countries. Also, research beyond the European countries is needed for the same reason. The consequences of Brexit

put even more strain on the linguistic landscape of British and European higher education and therefore, more research into language and literacy skills of international students not only in the UK, but elsewhere is urgently needed.

Chapter 7: Conclusions

This longitudinal study investigated language and literacy skills in undergraduate students at the start of university and one year later in a context of a heterogeneous, in terms of their first languages, student population. The primary aim of this investigation was to better understand language and literacy skills in students who pursue their degrees in their first language, and those for whom English is a foreign language. The second aim of this study was to find out if large and significant differences found between British home students and international students from China generalise to students from Europe, a group that is very different in many aspects, to validate previous findings.

This study found that rapid internationalisation leads to increase in linguistic diversity among students at a UK university. This is due to large and significant differences between British home students and the whole group of international students. Those who study in their native language had stronger language skills than the international cohort in all measures investigated at the start of Year one at university. Large and significant differences between British home students and EFL students with Chinese L1s have confirmed previous findings. However, these differences cannot be generalised to all student populations. This is because the difference in language and literacy kills between British home students and the EFL group with European L1s was not as great as the difference between British home students and international students from China.

EFL students speaking European L1s were much closer in their performance to the group of ENS at the start of Year one and managed to close the initial gap that existed between them and the group of British home students by the beginning of the second year. The group of EFL with Chinese L1, on the other hand, remained significantly different not only from the ENS group but also the other international group at both time points. The development of vocabulary seems to be the most challenging learning task for all students, as it was not possible, even for students speaking European L1s, to catch up with the group of ENS in this measure over the course of one year. However, the knowledge of vocabulary they arrived with allowed them to learn at a rate comparable to yearly growth in native speakers of English.

The findings are not likely to be explained by differences in cognitive abilities as no significant differences were found across all three groups in this dimension. The first language spoken can, to some extent, explain the findings as Chinese is a language very different from English which puts additional learning burden on those for whom it is a first language. However, there are other possible reasons as Chinese L1 students arrived with significantly weaker skills as indicated by the results obtained in standard English proficiency tests. They seemed to use English less than European L1 students while at university. The differences between the two international groups may also stem from cultural and educational distance from the UK with Chinese L1 students having a steeper learning curve when compared to European L1 students. It is also important to bear in mind that the study was conducted in unusual circumstances of lockdowns and social distancing, and this study should be ideally replicated in typical conditions for further validation.

The findings suggest that international students cannot be treated as a homogeneous group as variation exists across different cohorts. Chinese L1 students is a group that may benefit the most from language support while at university. The language provision should aim further development of vocabulary, reading and writing skills, because these are the most important linguistic factors that can predict academic achievement, and at the same time, skills than Chinese L1 students struggle with the most. They are likely to be the reason behind the attainment differences found between home and international students.

All international students, and especially those from China, are vital part of UK economy. To sustain economic activity in this sector, we need to be sure that all international students arrive with suitable level of proficiency in English, and those deemed to be at risk need to be taken care of. More research into predictors of academic success identifying potential problem areas is needed. This in turn would feed into pedagogical and policy-related decisions and interventions, help to improve the support offered to international students, and enable a level playing field regarding academic attainment.

Appendix A: Summary tables

Table A.1 Summary of the planned contrasts and post hoc tests at T1

Measure	Planned Contrast 1 (ENS vs EFL)	Planned Contrast 2 (European L1s vs Chinese L1s)	Post-hoc test results
VST raw	.000*	.000*	ENS>Eu (p < .001*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
VST response time	.000*	.000*	ENS>Eu (p < .001*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Grammar	.000*	.000*	ENS>Eu (p = .020*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Sight word efficiency	.000*	.000*	ENS=Eu (p = .978) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Reading rate	.000*	.000*	ENS>Eu (p < .001*) ENS>Ch (p < .001*) Eu>Ch (p = .002*)
Timed reading comprehension	.000*	.000*	ENS>Eu (p < .001*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Untimed reading comprehension	.000*	.000*	ENS=Eu (p = .105) ENS>Ch (p < .001*) Eu>Ch (p = .032*)
Number of words	.000*	.000*	ENS=Eu (p = .167) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Writing rate	.000*	.000*	ENS>Eu (p = .027*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Summarisation	.000*	.000*	ENS>Eu (p = .010*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Spelling	.000*	.001*	ENS=Eu (p = .093) ENS<Ch (p < .001*) Eu<Ch (p = .003*)
Elision	.010*	.000*	ENS=Eu (p = .983) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
RAN digits	.007*	.355	ENS=Eu (p = .065) ENS>Ch (p = .029*) Eu=Ch (p = .661)

Note: ENS = English native speakers, Eu = EFL speaking European L1s, Ch = EFL speaking Chinese L1s

Table A.2 Summary statistics for all measures obtained at T1

Measure	Group	M	SD	95% CI
VST raw	ENS	118.75	7.67	116.75—120.75
	EFL European L1s	103.85	10.89	101.01—106.69
	EFL Chinese L1s	73.76	14.87	69.85—77.67
VST response time	ENS	5,018	1,052	4,743—5,292
	EFL European L1s	6,790	1,669	6,355—7,225
	EFL Chinese L1s	9,142	2,851	8,392—9,892
Grammar	ENS	91.31	2.96	90.53—92.08
	EFL European L1s	89.67	3.87	88.67—90.67
	EFL Chinese L1s	80.93	6.73	79.16—82.70
Sight word efficiency	ENS	90.49	9.41	88.04—92.94
	EFL European L1s	90.82	8.21	88.70—92.94
	EFL Chinese L1s	77.95	9.22	75.52—80.37
Reading rate	ENS	222.66	77.47	202.47—242.85
	EFL European L1s	174.38	66.19	157.28—191.48
	EFL Chinese L1s	132.96	51.49	119.43—146.50
Timed reading comprehension	ENS	25.42	5.99	23.86—26.98
	EFL European L1s	20.37	6.85	18.60—22.14
	EFL Chinese L1s	12.22	5.82	10.69—13.75
Untimed reading comprehension	ENS	0.82	0.10	0.80—0.85
	EFL European L1s	0.78	0.12	0.75—0.81
	EFL Chinese L1s	0.72	0.14	0.68—0.76
Number of words	ENS	161.05	33.19	152.40—169.70
	EFL European L1s	148.85	39.68	138.60—159.10
	EFL Chinese L1s	114.67	33.28	105.92—123.42
Writing rate	ENS	18.52	3.80	17.53—19.51
	EFL European L1s	16.69	3.81	15.71—17.68
	EFL Chinese L1s	13.20	3.82	12.19—14.20
Summarisation	ENS	8.37	2.65	7.68—9.06
	EFL European L1s	6.97	2.54	6.31—7.62
	EFL Chinese L1s	4.31	2.13	3.75—4.87
Spelling	ENS	0.94	0.79	0.73—1.15
	EFL European L1s	1.36	1.31	1.02—1.69
	EFL Chinese L1s	2.40	2.00	1.87—2.93
Elision	ENS	17.56	2.20	16.99—18.13
	EFL European L1s	17.63	2.34	17.03—18.24
	EFL Chinese L1s	15.50	2.66	14.80—16.20
RAN digits	ENS	2.79	0.50	2.66—2.92
	EFL European L1s	2.58	0.47	2.47—2.71
	EFL Chinese L1s	2.49	0.70	2.30—2.68
Matrix reasoning	ENS	20.66	3.14	19.84—21.48
	EFL European L1s	19.98	2.89	19.23—20.74
	EFL Chinese L1s	21.02	2.24	20.43—21.61
Digit span backward	ENS	4.93	1.31	4.59—5.27
	EFL European L1s	4.71	0.99	4.45—4.96
	EFL Chinese L1s	5.07	1.32	4.72—5.42

Table A.3 Time 1 normality test results for all measures

Task	Group	Kolmogorov-Smirnov Sig. (*p*)	Shapiro-Wilk Sig. (*p*)
VST raw	ENS	.200	.243
	EFL European L1s	.200	.678
	EFL Chinese L1s	.000	.000
VST response time	ENS	.200	.415
	EFL European L1s	.200	.150
	EFL Chinese L1s	.200	.024
Grammar	ENS	.001	.001
	EFL European L1s	.005	.001
	EFL Chinese L1s	.200	.001
Sight word efficiency	ENS	.200	.040
	EFL European L1s	.200	.243
	EFL Chinese L1s	.200	.425
Reading rate	ENS	.019	.000
	EFL European L1s	.003	.004
	EFL Chinese L1s	.200	.081
Timed reading comprehension	ENS	.044	.036
	EFL European L1s	.002	.032
	EFL Chinese L1s	.009	.002
Untimed reading comprehension	ENS	.200	.000
	EFL European L1s	.200	.103
	EFL Chinese L1s	.200	.577
Number of words	ENS	.200	.389
	EFL European L1s	.200	.135
	EFL Chinese L1s	.200	.549
Writing rate	ENS	.200	.492
	EFL European L1s	.200	.636
	EFL Chinese L1s	.200	.901
Summarisation	ENS	.167	.273
	EFL European L1s	.035	.149
	EFL Chinese L1s	.000	.003
Spelling	ENS	.001	.000
	EFL European L1s	.000	.000
	EFL Chinese L1s	.004	.000
Elision	ENS	.000	.000
	EFL European L1s	.000	.000
	EFL Chinese L1s	.083	.132
RAN digits	ENS	.200	.546
	EFL European L1s	.002	.050
	EFL Chinese L1s	.001	.001
Matrix reasoning	ENS	.000	.000
	EFL European L1s	.200	.038
	EFL Chinese L1s	.002	.002
Digit span backwards	ENS	.000	.011
	EFL European L1s	.000	.008
	EFL Chinese L1s	.000	.000

Table A.4 Time 2 normality test results for all measures

Measure	Group	Kolmogorov-Smirnoff Sig. (*p*)	Shapiro-Wilk Sig. (*p*)
VST raw	ENS	.200	.067
	EFL European L1s	.041	.071
	EFL Chinese L1s	.024	.013
VST response time	ENS	.200	.521
	EFL European L1s	.200	.832
	EFL Chinese L1s	.012	.005
Grammar	ENS	.002	.010
	EFL European L1s	.004	.001
	EFL Chinese L1s	.200	.227
Sight word efficiency	ENS	.200	.027
	EFL European L1s	.200	.200
	EFL Chinese L1s	.200	.782
Reading rate	ENS	.200	.296
	EFL European L1s	.200	.001
	EFL Chinese L1s	.007	.001
Timed reading comprehension	ENS	.167	.029
	EFL European L1s	.200	.627
	EFL Chinese L1s	.200	.180
Untimed reading comprehension	ENS	.200	.232
	EFL European L1s	.197	.223
	EFL Chinese L1s	.011	.001
Number of words	ENS	.200	.984
	EFL European L1s	.200	.537
	EFL Chinese L1s	.053	.242
Writing rate	ENS	.200	.435
	EFL European L1s	.200	.804
	EFL Chinese L1s	.200	.226
Summarisation	ENS	.001	.206
	EFL European L1s	.001	.001
	EFL Chinese L1s	.001	.005
Spelling	ENS	.024	.001
	EFL European L1s	.092	.001
	EFL Chinese L1s	.001	.001
Elision	ENS	.001	.001
	EFL European L1s	.001	.001
	EFL Chinese L1s	.195	.148
RAN digits	ENS	.156	.118
	EFL European L1s	.004	.020
	EFL Chinese L1s	.087	.332

Table A.5 Summary of mean scores obtained at T1 and T2

Measure	Group	T1	T2	Change	*t*-test	Cohen's *d*	*p*
VST raw	ENS	118.56	120.27	1.71	-2.76	-.399	.004
	EFL European L1s	104.55	108.82	4.27	-5.70	-.815	.000
	EFL Chinese L1s	73.69	77.55	3.86	-3.41	-.487	.000
VST response time	ENS	5,119	4,599	519	4.13	.597	.000
	EFL European L1s	6,744	6,017	727	4.34	.620	.000
	EFL Chinese L1s	8,959	8,046	912	3.45	.493	.000
Grammar	ENS	91.44	91.50	0.06	-0.16	-.023	.437
	EFL European L1s	89.76	90.58	0.82	-1.76	-.249	.042
	EFL Chinese L1s	80.52	81.29	0.78	-0.97	-.140	.169
Sight word efficiency	ENS	90.46	91.96	1.50	-1.70	-.246	.047
	EFL European L1s	90.56	91.76	1.20	-1.70	-.240	.048
	EFL Chinese L1s	78.10	80.48	2.38	-2.93	-.423	.003
Reading rate	ENS	218.10	197.79	-20.31	2.51	.362	.008
	EFL European L1s	174.18	179.86	5.68	-0.84	-.119	.201
	EFL Chinese L1s	135.07	127.22	-7.82	0.94	.136	.176
Timed reading comprehension	ENS	25.29	25.56	0.27	-0.39	-.057	.348
	EFL European L1s	20.76	22.44	1.68	-2.46	-.348	.009
	EFL Chinese L1s	12.33	13.27	0.94	-1.61	-.232	.057
Untimed reading comprehension	ENS	.83	.82	.01	0.67	.097	.253
	EFL European L1s	.79	.78	.01	0.42	.059	.340
	EFL Chinese L1s	.70	.67	.03	1.15	.165	.129
Number of words	ENS	164.56	159.48	-5.08	1.16	.168	.126
	EFL European L1s	145.40	150.46	5.06	-1.02	-.143	.158
	EFL Chinese L1s	117.71	116.44	-1.27	0.29	.041	.388
Writing rate	ENS	18.72	19.31	.589	-1.25	-.180	.109
	EFL European L1s	16.56	18.39	1.83	-3.68	-.521	.000
	EFL Chinese L1s	13.61	14.16	0.55	-1.10	-.159	.138
Summarisation	ENS	8.31	8.79	0.48	-1.27	-.184	.105
	EFL European L1s	6.90	7.70	0.80	-1.83	-.258	.037
	EFL Chinese L1s	4.31	5.08	0.77	-2.68	-.387	.005
Spelling	ENS	0.90	1.00	0.10	-0.67	-.097	.253
	EFL European L1s	1.43	1.58	0.14	-0.76	-.108	.226
	EFL Chinese L1s	2.27	3.46	1.19	-3.51	-.507	.000
Elision	ENS	17.56	17.90	0.33	-1.44	-.208	.078
	EFL European L1s	17.62	17.92	0.30	-0.97	-.138	.167
	EFL Chinese L1s	15.56	15.96	0.40	-1.26	-.182	.107
RAN digits	ENS	2.79	3.01	0.22	-3.45	-.497	.000
	EFL European L1s	2.56	2.74	0.18	-3.39	-.479	.000
	EFL Chinese L1s	2.56	2.59	0.03	-0.27	-.040	.393

Table A.6 Summary of significance test results obtained in mixed design ANOVA

Measure	Main effect of Time	Main effect of Group	Time * Group Interaction	Main Effect of Group post hoc tests
VST raw	$p < .001$*	$p < .001$*	$p = .090$	ENS>Eu ($p = .000$*) ENS>Ch ($p = .000$*) Eu>Ch ($p = .000$*)
VST Response time	$p < .001$*	$p < .001$*	$p = .369$	ENS>Eu ($p = .000$*) ENS>Ch ($p = .000$*) Eu>Ch ($p = .000$*)
Grammar	$p = .099$	$p < .001$*	$p = .584$	ENS=Eu ($p = .066$) ENS>Ch ($p < .001$*) Eu>Ch (p < .001*)
Sight word efficiency	$p < .001$*	$p < .001$*	$p = .560$	ENS=Eu ($p = .999$) ENS>Ch ($p < .001$*) Eu>Ch ($p < .001$*)
Reading rate	$p = .095$	$p < .001$*	$p = .610$	ENS>Eu ($p = .049$*) ENS>Ch ($p < .001$*) Eu>Ch ($p = .001$*)
Timed reading comprehension	$p = .012$*	$p < .001$*	$p = .314$	ENS>Eu ($p = .007$*) ENS>Ch ($p < .001$*) Eu>Ch ($p < .001$*)
Untimed reading comprehension	$p = .181$	$p < .001$*	$p = .727$	ENS=Eu ($p = .140$) ENS>Ch ($p < .001$*) Eu>Ch ($p < .001$*)
Number of words	$p = .872$	$p < .001$*	*$p = .293$*	ENS=Eu ($p = .119$) ENS>Ch ($p < .001$*) Eu>Ch ($p < .001$*)
Writing rate	p < .001*	$p < .001$*	*$p = .107$*	ENS=Eu ($p = .063$) ENS>Ch ($p < .001$*) Eu>Ch ($p < .001$*)
Summarisation	$p = .002$*	$p < .001$*	$p = .800$	ENS>Eu ($p = .025$*) ENS>Ch ($p < .001$*) Eu>Ch ($p < .001$*)
Spelling	$p = .001$*	$p < .001$*	$p = .002$*	ENS=Eu ($p = .135$) ENS>Ch ($p < .001$*) Eu>Ch ($p < .001$*)
Elision	$p = .041$*	$p < .001$*	$p = .972$	ENS=Eu ($p = .994$) ENS>Ch ($p = .000$*) Eu>Ch ($p < .001$*)
RAN digits	$p = .002$*	$p = .005$*	$p = .196$	ENS>Eu ($p = .046$*) ENS>Ch ($p = .008$*) Eu=Ch ($p = .675$)

Note: ENS = English native speaker, Eu = EFL speaking European L1s, Ch = EFL speaking Chinese L1s

Table A.7 Summary of post hoc tests at both time points

Task	Post hoc tests T1[a]	Post hoc tests T2
VST raw	ENS>Eu (p < .001*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)	ENS>Eu (p < .001*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
VST response time	ENS>Eu (p < .001*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)	ENS>Eu (p < .001*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Grammar	ENS>Eu (p = .043*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)	ENS=Eu (p = .291) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Sight word efficiency	ENS=Eu (p = .998) ENS>Ch (p < .001*) Eu>Ch (p < .001*)	ENS=Eu (p = .993) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Reading rate	ENS>Eu (p = .010*) ENS>Ch (p < .001*) Eu>Ch (p = .005*)	ENS=Eu (p = .469) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Timed reading comprehension	ENS>Eu (p = .002*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)	ENS=Eu (p = .063) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Untimed reading comprehension	ENS=Eu (p = .186) ENS>Ch (p < .001*) Eu>Ch (p = .007*)	ENS=Eu (p = .296) ENS>Ch (p < .001*) Eu>Ch (p = .005*)
Number of words	ENS>Eu (p = .035*) ENS>Ch (p < .001*) Eu>Ch (p = .001*)	ENS=Eu (p = .501) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Writing rate	ENS>Eu (p = .018*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)	ENS=Eu (p = .422) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Summarisation	ENS>Eu (p = .028*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)	ENS=Eu (p = .135) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Spelling	ENS=Eu (p = .050) ENS<Ch (p < .001*) Eu<Ch (p = .018*)	ENS>Eu (p = .035*) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
Elision	ENS=Eu (p = .992) ENS>Ch (p < .001*) Eu>Ch (p < .001*)	ENS=Eu (p = .998) ENS>Ch (p < .001*) Eu>Ch (p < .001*)
RAN digits	ENS=Eu (p = .072) ENS=Ch (p = .152) Eu=Ch (p = .988)	ENS=Eu (p = .068) ENS>Ch (p = .001*) Eu=Ch (p = .230)

[a]***Note.*** This applies to a subset of T1 participants tested across both time points

Table A.8 Descriptive statistics for all measures and gains across two time points in a subset of participants tested at T1 and T2

Measure	ENS			EFL					
				Europeans L1s			Chinese L1s		
	N	M	SD	N	M	SD	N	M	SD
Vocabulary									
VST estimated T1	48	16,316	3,988	49	11,542	2,881	49	7,446	1,705
VST estimated T2	48	17,085	3,460	49	12,806	3,493	49	7,857	1,821
VST estimated T2–T1	48	769	2,196	49	1,263	1,571	49	410	938
VST raw T1	48	118.56	7.79	49	104.55	11.06	49	73.69	14.70
VST raw T2	48	120.27	6.40	49	108.82	11.30	49	77.55	14.96
VST raw T2–T1	48	1.71	4.28	49	4.27	5.24	49	3.86	7.92
VST response time T1	48	5,119	1,072	49	6,744	1,725	49	8959	2,822
VST response time T2	48	4,599	1,049	49	6,017	1,525	49	8046	2,681
VST response time T2–T1	48	-519	871	49	-727	1,172	49	-912	1,852
Grammar									
Grammar T1	48	91.44	2.80	50	89.76	3.93	48	80.52	6.66
Grammar T2	48	91.50	2.48	50	90.58	3.48	48	81.29	6.46
Grammar T2 – T1	48	0.06	2.72	50	0.82	3.29	48	0.77	5.51
Word/Text reading									
Sight word efficiency T1	48	90.46	9.24	50	90.56	8.68	48	78.10	9.16
Sight word efficiency T2	48	91.96	9.27	50	91.76	8.04	48	80.48	9.92
Sight word efficiency T2–T1	48	1.50	6.10	50	1.20	5.01	48	2.38	5.61
Reading rate T1	48	218.10	77.71	50	174.18	67.95	48	135.07	53.08
Reading rate T2	48	197.79	68.25	50	179.86	81.99	48	127.22	56.93
Reading rate T2–T1	48	-20.31	56.12	50	5.68	47.57	48	-7.86	57.81
Timed reading comprehension T1	48	25.29	5.95	50	20.76	6.79	48	12.33	6.10
Timed reading comprehension T2	48	25.56	6.75	50	22.44	6.79	48	13.27	6.47
Timed reading comprehension T2–T1	48	0.27	4.79	50	1.68	4.83	48	0.94	4.04

Untimed reading comprehension T1	48	.83	.10	50	.79	.12	48	.70	.14
Untimed reading comprehension T2	48	.82	.11	50	.78	.13	48	.67	.19
Untimed reading comprehension T2–T1	48	-.01	.11	50	-.01	.13	48	-.03	.17
Text writing									
Number of words T1	48	164.56	34.47	50	145.40	40.49	48	117.71	30.26
Number of words T2	48	159.48	38.88	50	150.46	40.46	48	116.44	37.71
Number of words T2–T1	48	-5.08	30.34	50	5.06	35.26	48	-1.27	30.82
Writing rate T1	48	18.72	3.76	50	16.56	3.98	48	13.61	3.36
Writing rate T2	48	19.31	3.70	50	18.39	3.55	48	14.16	3.97
Writing rate T2–T1	48	0.59	3.26	50	1.83	3.52	48	0.55	3.44
Summarization T1	48	8.31	2.68	50	6.90	2.68	48	4.31	2.10
Summarization T2	48	8.79	2.45	50	7.70	3.11	48	5.08	2.44
Summarization T2–T1	48	0.48	2.61	50	0.80	3.10	48	0.77	1.99
Spelling error T1	48	0.90	0.76	49	1.44	1.39	48	2.27	1.74
Spelling error T2	48	0.99	0.91	49	1.58	1.33	48	3.46	2.91
Spelling error T2–T1	48	0.09	1.01	49	0.14	1.33	48	1.19	2.35
Phonological processing									
Elision T1	48	17.56	2.15	50	17.62	2.47	48	15.56	2.76
Elision T2	48	17.90	1.93	50	17.92	1.76	48	15.96	2.49
Elision T2–T1	48	.34	1.60	50	0.30	2.18	48	0.40	2.18
RAN digits T1	48	2.79	0.52	50	2.56	0.48	47	2.56	0.75
RAN digits T2	48	3.01	0.64	50	2.74	0.49	47	2.59	0.44
RAN digits T2–T1	48	0.22	0.44	50	0.18	0.38	47	0.03	0.74
Cognitive Abilities									
Matrix reasoning T1	59	20.66	3.14	59	19.98	2.90	58	21.02	2.24
Digit span backwards T1	59	4.93	1.31	60	4.71	0.99	58	5.07	1.32

Pre-study Survey

Title of the study: Language and literacy skills in university students and their success in higher education.

Study Screening Survey

Thank you for your interest in our study. We are interested in how language develops in tertiary education and how it relates to academic success. This is a very brief survey, which will take just a few minutes to complete. The aim is to check whether your profile matches the study criteria.

What you will be asked to do?

You will be asked a few questions about yourself (e.g., the languages you speak). If your profile matches the criteria, we will send you further information about the study and invite you to take part. You will need to leave your name and email address at the end of this survey so that we can contact you. If your profile does not match the study criteria, you will know by the end of this survey and will not be asked to leave your contact details.

What will happen to the information you share with us?

We will keep all the information about participants in this study confidential and in accordance with the GDPR. All data will be stored securely on a password-protected computer. The data you provide will be stored by a randomly assigned code number. Any information that identifies you will be stored separately from the data. Only I (Justyna Mackiewicz) will have access to that information. The personal identifiers will only be kept until the data collection for the study is completed. The fully anonymised data files may be kept indefinitely and may be shared with other researchers or used in future research.

Participation is voluntary

If you decide to take part, you will be asked to complete a consent form. If you change your mind at any point during the study, you will be able to withdraw your participation without having to provide a reason.

Questions or concerns

This research has been approved by the Department of Education, University of York Ethics Committee. If you have any questions or complaints about this research, please contact the researcher through the email: jm1730@york.ac.uk or the Ethics Committee via education-research-administrator@york.ac.uk. If you are still dissatisfied, please contact the University's Data Protection Officer at dataprotection@york.ac.uk. If you have any questions, please feel free to contact me.

Justyna Mackiewicz, jm1730@york.ac.uk.

CONSENT FORM

Title of the study: *Language and literacy skills in university students and their success in higher education.*

I confirm that I was given information about this pre-study survey and had the opportunity to ask questions.

I understand that participation in this study is voluntary.

I understand that I need to answer all questions truthfully for the researcher to determine whether my profile matches the study criteria.

I understand that if I don't answer all questions, I will not be able to participate in the study.

I understand that if my profile matches the study criteria, I will need to provide my name and email address for the researchers to contact me. The researchers will only use this information to send me more information about the study.

I understand that any information about me will be held confidentially in accordance with the University regulations and GDPR.

Please tick 'YES' below if you are happy to take part in this pre-study survey.

○ **YES** (1)

Q1. I am currently a first-year undergraduate student at the University of York.

○ Yes (1)

○ No (2)

Skip To: Q11 If I am currently a first-year undergraduate student at the University of York. = No

Skip To: Q2 If I am currently a first-year undergraduate student at the University of York. = Yes

Q2. I am going to study in York for at least 3 years.

○ True (1)

○ False (2)

Skip To: Q3 If I am going to study in York for at least 3 years. = True

Skip To: Q11 If I am going to study in York for at least 3 years. = False

Q3. I have not studied at a university before. This is my first undergraduate degree.

○ True; I have not studied at a university before. This is my first undergraduate degree. (1)

○ False; I have studied for another undergraduate degree before. (2)

Skip To: Q11 If I have not studied at a university before. This is my first undergraduate degree. = False; I have studied for another undergraduate degree before.

Q4. I do not have dyslexia or a formally diagnosed condition that may affect my vision, language, reading, writing or hearing. (Don't worry if you wear glasses or contact lenses - that's fine and doesn't fall under 'formally diagnosed condition affecting vision').

○ True; I do not have dyslexia or any condition that may affect my vision, language, reading, writing or hearing. (1)

○ False; I have a condition that can affect my vision, hearing, language or literacy. (2)

Skip To: Q11 If I do not have dyslexia or a formally diagnosed condition that may affect my vision, language, rea... = False; I have a condition that can affect my vision, hearing, language or literacy.

Now we would like to ask you a few questions about your language background.

Q5. In which department are you studying?

_____

Q6. Is your first language English? (First language = the first language you learnt at home)

○ Yes, English is my first language (1)

○ No, English is not my first language (2)

Skip To: Q11 If Is your first language English? (First language = the first language you learnt at home) = Yes, English is my first language

Q7. Are you an international student? (= does not normally live in the UK)

○ Yes (1)

○ No (2)

Skip To: Q8 If Are you an international student? (= does not normally live in the UK) = Yes

Skip To: Q9 If Are you an international student? (= does not normally live in the UK) = No

Q8. Are you an EU student?

○ Yes (1)

○ No (2)

Skip To: Q12 If Are you an EU student? = No

Skip To: Q8 If Are you an EU student? = Yes

Q9. Which of the following describes you best?

○ My first language is Chinese (1)

○ My first language is one of the European languages (2)

○ None of the above (3)

Skip To: Q10 If Which of the following describes you best = My first language is Chinese

Skip To: Q10 If Which of the following describes you best = My first language is one of the European languages

Skip To: Q11 If Which of the following describes you best = None of the above

Q10. Thank you for submitting your answers. I will check them and get back to you to confirm whether your profile matches all the study requirements. If you are happy for me to contact you, please leave your contact details below.

After that press 'next' to submit the form.

○ **Your name:** (1) _____

○ **Your email:** (2) _____

Skip To: End of Survey If Condition: Your name: Is Displayed. Skip To: End of Survey.

Skip To: End of Survey If Condition: Your email: Is Displayed. Skip To: End of Survey.

Q11. Thank you again for your interest in our study and taking the time to complete this survey. Unfortunately, your profile does not match the study criteria. If you have friends who meet the study criteria, feel free to let them know about this study. Press the 'next' button now to submit your responses.

Appendix C: T1 information page and consent form

Participant Information Page

Title of the study: Language and literacy skills in university students and their success in higher education.

Researcher: Ms. Justyna Mackiewicz, University of York.

Supervisor: Dr Danijela Trenkic, University of York.

This study is a part of my PhD research project that investigates language and literacy skills in university students from a variety of linguistic backgrounds. I am interested in how their language develops over a course of an academic year and how this affects academic performance. Your participation will help us to understand better some language related issues at the tertiary level and the findings will benefit those who start university in the future.

Before agreeing to take part, please read this information carefully and let me know if anything is unclear or you would like further information. Please also read and keep for your records the enclosed General Data Protection Regulation (GDPR) document. If you have further question after the study, you can email the researcher, Justyna Mackiewicz, at: jm1730@york.ac.uk.

What you will be asked to do?

The data collection for the present study will be held at two points in time: at the beginning of your first and second year at the university. This means that in order to take part in the study, you need to be physically present on the University of York campus and willing to participate in both testing sessions. In year one we will ask you to complete a number of language and memory tasks. You will attend 2 sessions with the researcher, each lasting about 1 hour. This will be repeated in year two. As a thank-you for your participation, you will receive £15 each year upon completing all tasks.

Due to the design of the study, we will need to know all your course marks. You need to be happy to allow the university to share your module marks with the researcher. As the information on your course marks is essential for the study, you cannot take part in the study if you do not wish to consent access to your marks. Please note that an ID code will be used to store your data and no one apart from the researcher will be able to link marks to individual names.

What will happen to the information you share with us?

We will keep all the information about participants in this study confidential. All data, including your marks and any tests you take, will be stored securely. You will be randomly assigned an ID number, and this will be the only form of identification that will be included on any database and paper-based tasks used in this study. Any information that identifies you will be stored separately from the data. The encrypted data files and key to the ID codes will be stored in separate password-protected folders on university secure servers. The personal identifiers will only be kept until the data collection for the full study is completed. The anonymised data files may be kept indefinitely, shared with other researchers, or used in future research.

After the transfer of data to electronic data files, any hard copies of data will be permanently destroyed using University's confidential waste disposal services. The consent forms will be kept securely until the completion of the study and then will be destroyed. Your name will never be included in any publications (e.g. conference presentations, journal articles) based on this study.

Do I have the right to withdraw?

YES. Your participation in this study is completely voluntary. You are free to stop your participation at any point during the data collection without giving a reason. You also have 20 days after your last participation in the study to request the withdrawal of your data by writing to jm1730@york.ac.uk.

Questions or concerns

This research has been approved by the Department of Education, University of York Ethics Committee. If you have any questions or complaints about this research please contact the researcher through the email: jm1730@york.ac.uk or the Ethics Committee via education-research-administrator@york.ac.uk. If you are still dissatisfied, please contact the University's Data Protection Officer at dataprotection@york.ac.uk .

Please save this information for your own records.

Thank you very much for taking the time to read this information.

Yours Sincerely

Justyna Mackiewicz

Language and literacy skills in university students and their success in higher education

Consent Form

Please tick each box if you are happy to take part in this research.

I confirm that I have read and understood the information given to me about the above named research project and I understand that this will involve me taking part as described above.	
I agree for the University to share my course marks with the researcher for the purpose of this research study.	
I understand that participation in this study is voluntary.	
I understand that my data will not be identifiable and the data may be used in publications, presentations and online.	
I confirm that I have read the information about GDPR.	

NAME: ……

STUDENT NUMBER: ……

DATE: ……

SIGNATURE: ……

Appendix D: T2 information page and consent form

Participant Information Page

Title of the study: Language and literacy skills in university students and their success in higher education.

Researcher: Ms. Justyna Mackiewicz, University of York.

Supervisor: Dr Danijela Trenkic, University of York.

Thank you for taking part in the first round of data collection for this study last year, and for expressing your willingness to come back for the second round this year. Because of Covid 19, data collection this year will be conducted entirely on-line. Please read the information below carefully and let me know if anything is unclear or you would like further information. If you have further question after the study, you can email me (Justyna Mackiewicz) at jm1730@york.ac.uk.

What is the study about?
The study is a part of my PhD research project that investigates language and literacy skills in university students from a variety of linguistic backgrounds. I am interested in how their language develops over a course of an academic year and how this affects academic performance. Your participation will help us to understand better some language related issues at the tertiary level and the findings will benefit those who start university in the future.

What will you be asked to do this time?
In year one we asked you to complete a number of language and memory tasks. You attended 2 sessions with the researcher, each lasting about 1 hour. This design will be repeated this year, but both sessions will be held online via Zoom. This means that you do not have to be physically present on the campus. As previously agreed, you will receive £15 as a thank-you after completing both research sessions this year. However, because of the social distancing requirements, you will receive £15 as an Amazon voucher, sent to you by email.

What will happen to the information you share with us?
We will continue to keep all the information about participants in this study confidential. All data is stored securely. You have been randomly assigned an ID number, and this is the only form of identification that will be included on any database. Any information that identifies you will be stored separately from the data. The encrypted data files and key to the ID codes

are stored in separate password-protected folders on university secure servers. The personal identifiers will only be kept until the data collection for the full study is completed. The anonymised data files may be kept indefinitely, shared with other researchers, or used in future research. Your name will never be included in any publications (e.g., conference presentations, journal articles) based on this study.

Processing of your data – GDPR statement

Information that you provide will be treated confidentially and shared on a need-to-know basis only. The University of York is committed to the principle of data protection by design and default and will collect the minimum amount of data necessary for the project. In line with our charter which states that we advance learning and knowledge by teaching and research, we process personal data for research purposes under Article 6(1) (e) of the GDPR: *Processing is necessary for the performance of a task carried out in the public interest.* Special category data is processed under Article 9 (2) (j): *Processing is necessary for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes.*

Do I have the right to withdraw?

YES. Your participation in this study is completely voluntary. You are free to stop your participation at any point during the data collection without giving a reason. You also have 20 days after your last participation in the study to request the withdrawal of your data by writing to jm1730@york.ac.uk.

Questions or concerns

This research has been approved by the Department of Education, University of York Ethics Committee. If you have any questions or complaints about this research please contact the researcher through the email: jm1730@york.ac.uk or the Ethics Committee via education-research-administrator@york.ac.uk. If you are still dissatisfied, please contact the University's Data Protection Officer at dataprotection@york.ac.uk.

Thank you very much for taking the time to read this information.
(Please move to the next page)

Language and literacy skills in university students and their success in higher education

Consent Form

I confirm that I was given information about this study and I understand what my participations will involve.

I understand that participation in this study is voluntary and I am happy to take part.

I understand that I can stop participating at any point during data collection and can withdraw my data up to 20 days after data collection by writing to Justyna Mackiewicz.

I understand that any information about me will be held confidentially in accordance with the University regulations and GDPR.

If you agree with all of the above statements and wish to proceed with the study, please type your name here:

_____

Background questionnaire

Participant ID:

_____

Language Background Questionnaire

Q1. How old are you?

_____

Q2. What is your gender/how do you identify?

○ Male (1)

○ Female (2)

○ Other (3)

Q3. In which country were you born?

_____

Q4. How long have you been living in the UK? (state clearly the number of months or years)

_____

Q5. List all the languages other than English that you can understand, speak, read or write. If none, then state 'none'.

_____

Q6. What is the highest level of education for your mother/primary caregiver?

- ○ Some or no secondary education (1)
- ○ Secondary school education (2)
- ○ Post-secondary education with vocational training (3)
- ○ University degree (4)
- ○ Post-graduate degree or professional education (5)
- ○ Don’t know/ not applicable (6)

Q7. What is the highest level of education for your father/ primary caregiver.

- ○ Some or no secondary education (1)
- ○ Secondary school education (2)
- ○ Post-secondary education with vocational training (3)
- ○ University degree (4)
- ○ Post-graduate degree or professional education (5)
- ○ Don’t know/ not applicable (6)

Q8. Please indicate the language(s) spoken by your mother/ primary caregiver.

- ○ First language (2) _____
- ○ Second language (if applicable) (3) _____
- ○ Other additional languages (if applicable) (8) _____
- ○ Do not know (9) _____

Q9. Please indicate the language(s) spoken by your father/primary caregiver.

○ First language (1) _____

○ Second language (if applicable) (2) _____

○ Other additional languages (if applicable) (3) _____

○ Do not know (7) _____

Q10. Is English your first language?

○ Yes (1)

○ No (2)

Skip To: End of Block If Is English your first language? = Yes

Skip To: Q11 If Is English your first language? = No

Q11. What is your first language or languages (those that you have learnt from birth)?

_____

Q12. Did anyone from your immediate family (for example, mother, father, grandparents, siblings) speak to you in English when you were growing up at home?

○ Yes (1)

○ No (2)

Skip To: Q13 If Did anyone from your immediate family (for example, mother, father, grandparents, siblings) speak... = Yes

Skip To: Q14 If Did anyone from your immediate family (for example, mother, father, grandparents, siblings) speak... = No

Q13. How often?

○ Daily (3)

○ Often (4)

○ Rarely (5)

Q14. How old were you when you first began to study English at school?

_____

Q15. Have you attended a school where you learned all subjects in English?

○ Yes (1)

○ No (2)

Skip To: Q16 If Have you attended a school where you learned all subjects in English? = Yes

Skip To: Q17 If Have you attended a school where you learned all subjects in English? = No

Q16. What is the level of the qualifications you obtained in this English speaking institution (e.g. GCSE, A levels, etc.)?

_____

Q17. Have you lived in an English-speaking country prior to starting university (e.g. England, the US, Canada, Australia, New Zealand)?

○ Yes (1)

○ No (2)

Skip To: Q18 If Have you lived in an English-speaking country prior to starting university (e.g. England, the US… = Yes

Skip To: Q19 If Have you lived in an English-speaking country prior to starting university (e.g. England, the US… = No

Q18. How long have you lived in an English-speaking country?

_____

Q19. How old were you when you arrived in the UK?

_____

Q20. Have you taken any standardised English language proficiency tests (e.g., TOEFL, IELTS)?

○ Yes (1)

○ No (2)

Skip To: Q21 If Have you taken any standardised English language proficiency tests (e.g., TOEFL, IELTS)? = Yes

Skip To: Q28 If Have you taken any standardised English language proficiency tests (e.g., TOEFL, IELTS)? = No

Q21. Which proficiency test did you take last before arriving at the university?

_____

Q22. When did you take this proficiency test? (Which year?)

_____

Q23. What was your overall score on this proficiency test?

_____

Q24. What was your score on the reading part?

_____

Q25. What was your score on the writing part?

_____

Q26. What was your score on the speaking part?

_____

Q27. What was your score on the listening part?

_____

Q28. Have you attended a pre-sessional English language programme before starting university? (a course offered by universities to international students to help them improve their English language skills before the start of their degree programmes)

○ Yes (1)

○ No (2)

Skip To: Q29 If Have you attended a pre-sessional English language programme before starting university? (a cours... = Yes

Skip To: End of Survey If Have you attended a pre-sessional English language programme before starting university? (a cours... = No

Q29. How long was your pre-sessional course? (e.g., 8 weeks, 6 months, 1 year etc.)

_____

Appendix F: Researcher's handbook

Researcher's Handbook

	Order of tasks	Timed?	Approximate length (min)
1	Vocabulary Size Test	No	50
2	Nelson-Denny	Yes	15
3	Which English grammar test	No	10
4	History of chocolate writing task	Yes	10
5	Matrix reasoning	To some extent	8
6	Digit span backward	No	2
7	Elision	No	3
8	Rapid automatic naming	No	1
9	Sight word efficiency	Yes	1
10	Main study survey	No	10

Vocabulary Size Test T1 instructions

Instructions for the in-person session conducted in the testing lab. They are presented on three separate PP slides to the whole group.

- Please switch off your mobile phone
- You will do a computer-based English vocabulary test
- You will be asked the meaning of 140 English words
- Each word will be shown in a sentence and you will be given 4 possible definitions to choose from
- You must not consult with anyone or use a dictionary
- The task takes around 30 minutes
- After you complete the last question (No. 140), something like an error message may appear on screen, with an option 'retry' and 'cancel'. **Press 'retry'**
- You will be asked a couple of language background questions
- When you are finished, you'll be given your result
- Once you get to the result page, you can close the browser, log off and leave quietly.
- Please go to the website indicated on the slip of paper I've given you, **Type the full address, including 'http' (or it won't work).**
- You will be asked for an access code, and participant ID They are on the same slip of paper
- Then follow the instructions on the screen
- If you have a question during the test, please raise your hand and I will come to see you

Vocabulary Size Test T2 instructions

These instructions are presented in a shared screen with the participant in the individual Zoom testing session at T2.

- You will be asked the meaning of **140** English words
- Each word will be shown in a sentence, and you will be given 4 possible definitions to choose from
- You must not use a dictionary or any other resources
- The task takes around **30 minutes**
- After you complete the last question (No. 140), something like an error message may appear on screen, with an option 'retry' and 'cancel'. **Press 'retry'**
- You will be asked a couple of language background questions
- When you are finished, you'll be given your result
- Once you get to the result page, you can close the browser.

Nelson-Denny reading comprehension test

Info for the researcher:

You will need a timer for this task. Set it to 15 minutes.

This is an online task. Open the relevant file from Qualtrics.

Fill in the Participant ID and move to the instruction page.

Tell the participant: **This is a reading comprehension task. Please read the instructions. If you have any questions, feel free to ask. Let me know when you are ready to start.**

When the participant is ready, tell them **You can start now** and at the same time start the timer (15 minutes). After 15 minutes, say **Time is up. Let's move to the next task**.

Note: Participants are not allowed to read the questions *before* reading the passage: they must read the passage first. But they can navigate backwards and forwards between the text and the questions once they've read the passage. Keep an eye on their behaviour and warn them if you notice that they read questions before reading the text. Keep a record of this. If participant does not answer all the questions before the end of the test then the data will not be recorded properly. So keep moving the pages by clicking the arrow at the end of the page until the end.

Instructions provided on the first page in Qualtrics:

There are seven reading passages in this reading task. Each is followed by multiple choice comprehension questions. You MUST complete reading each passage first before looking at the questions. Later, you may look back at the material you have read, but do not spend too long on any one question.

When you have completed the questions for one passage, go immediately to the next one.

You have 15 minutes to complete this reading task.

Continue working until you have answered all of the questions or until you are told to stop.

Your score is based on the number of *correct* responses. Since there is no penalty for incorrect answers, it is to your advantage to mark every question you read.

Please, do your best to try to answer all the given questions.

Which English grammar test

Info for the researcher:

This is an online task. It is not timed. It should take about 10 minutes to complete.

Open the relevant file from from Qualtrics. Fill in the Participant ID and move to the instruction page.

Tell the participant: **This is a grammar quiz. Please read the instructions. If you have any questions, feel free to ask. Let me know when you are ready to start.**

When the participant is ready, tell them: You can start now. When they have finished, check that the 'End of Survey' screen is on.

Instructions provided on the first page in Qualtrics:

In this quiz, you will decide which sentences are grammatical (correct) and which are not.

Do not worry about whether the sentence is formal or 'proper' or is what you learned at school. Scientists have discovered that many of the 'rules' taught in school are wrong anyway.

Focus on your gut instincts. Does the sentence sound correct, or does it sound like a mistake -- for instance, a mistake made by a young child or a second language learner?

Written precis: The history of chocolate

Info for the researcher: Prepare: stopwatch, YAA-R record form, YAA-R Written précis sheet.

Tell the participant: **You are now going to read another text but this time you will also write a summary of the main points. I am interested in both how quickly you can read a text silently, and how much of it you can remember afterwards. This means that when you are writing a summary, you will not be able to look at the text again. OK?**

Present the page with the history of chocolate and tell the participant: **Please read this text silently at your own pace. You should read straight through and not reread sections. While you read, I will be timing you. Please tell me when you have finished so I can stop the timer.**

Remove the page and tell the participant: **OK. You now have 10 minutes to write a summary of the main points. When I say stop, please stop writing; however, if you finish your summary before the time is up, please let me know**.

Info for the researcher: If the individual finishes before 10 minutes, please note the time on their sheet. Record time in seconds.

Matrix reasoning

Info for the researcher:

Check that you have correct participant ID at the top of the scoring sheet. Open the stimulus folder. Begin with the first two sample items to make sure participants understood the task

Show the sample A picture and explain: **Look at this picture and show me which of these** [*show the options below*] **goes here** [*show the box with question mark*]. **Note that when choosing your answer, you should look for a pattern going across** [*show with your finger*], **but never diagonally. Now show me which one goes here** [*show the box with question mark again*]?

If the correct answer is provided move to sample B and say: **This is another kind of problem. Here the boxes are only going across. Looking at the pattern going across, which one of these** [*show the options below*] **goes here** [*show the box with question mark*]?
If the person gets it wrong repeat the instructions, provide the correct answer and explain why until they understand the task. Start the task with item 1. Record answers on the form. Correct answers are highlighted on the scoring sheet. If the participant spends more than 30 seconds on an item, move them gently on: **Lets' try the next one!** Stop testing after 3 consecutive scores of 0.

Elision

Tell participant: **Let's play a word game now. Let me show you how to play it:**

Can you say *toothbrush* without *tooth*?

If the participant says *brush*, you say: **Good! Let's practise another example.**

If the answer is incorrect tell them the right answer and ask to do another example.

Go to the next practice example from the answer sheet.

Ask: **Say *airplane* without *plane*.**

If the participant says *air*, you say: **Good! Let's start the game now.**

When arriving at the single sound deletion: **The game changes now. I remove one sound rather than the whole word.**

Can you say *cup* without *k*?

If the participant says *up*, you say: **Good! Let's practice another example.**

If the answer is incorrect tell the right answer and ask to do another example.

Go to the next practice example from the answer sheet.

Ask: **Can you say *meet* without *t*?**

If the participant says *me*, you say: **Good! Let's come back to the game now.**

Digit span backward

Infor for the researcher:

Check that you have the correct participant ID at the top of the scoring sheet.

It is important to pace yourself for this task so that there is a 1 second interval between each digit. I tend to do this by saying 'one thousand' silently after each digit. If you are too fast or too slow it will impact on the participant performance.

Tell the participant: **I am now going to read lists of numbers. I will stop after each list. When I stop, I want you to say the numbers backwards (in the reverse order). For example, if I say 4- 9, you should say: 9- 4. I will read each sequence only once, so please, listen carefully. Are you ready?**

Start from the sample item. If the participant gets it wrong, provide feedback and give the second sample item. If they get it wrong, repeat the sample trial until they get it right, then move to item 1.

Do not provide any feedback, record answers on the form.

Stop testing after no sequence is recalled correctly within a group of 2 of the same length.

Rapid automatic naming

Info for the researcher:

Prepare: stopwatch, RAN digits test and practice items.

Display the practice items and tell the participant: **Please name the following digits aloud from left to right. Please, read as fast as you can but I still need to understand what you are saying. Let's begin with some practice items.**

Info for the researcher:

Check how long the individual takes to name the list of digits, noting errors.

Administer the practice items first. During the practice for digits, correct individual if he/she makes an error. Items are to be named from left to right. If they mispronounce words in the practice list, repeat. Read as fast as you can but I still need to understand what you are saying. Start timing once they start naming and stop timing when they name the last item. Record time in seconds.

Sight word efficiency

Prepare the practice items list, show it to the participant and say:

I want you to read a list of words as fast as you can without making errors. Let's start with this practice list. Begin at the top, read down the list as fast as you can. If you come to a word you cannot read, just skip it and go to the next word. Use your finger to help you keep your place if you want to.

Check if the participant has any questions and proceed to the real list of words. When the practice part is completed move to the main word list.

Tell the participant: **OK now you will read some longer lists of words. Read as many words as fast as you can without making errors, until I tell you to stop. OK? You will begin as soon as I turn over the card.**

Info for the researcher:

Turn over the practice word list so the test word list is exposed, and start timing. As they are reading, mark any words that are misread or skipped. If they hesitate for more than 3 seconds on a word, *say* **Move to the next one.** After 45 seconds, tell the participant to stop. Draw a line under the last word they read. If the person finishes all the words before the time is up, note the time taken them to read all the words.

Appendix G: History of Chocolate text

The History of Chocolate

Chocolate is now enjoyed all over the world but until the late sixteenth century it was only found in Central and South America. For years the indigenous people of Central and South America, the Aztecs and the Mayans, had been making chocolate from cacao beans and consuming it as a drink. Cacao beans carried great importance for both cultures: the Mayans considered them to be a gift from the gods and the Aztecs associated them with fertility. Both societies even used them as currency. Cacao beans were made into a chocolate drink by roasting them and adding water and chilli spice. The mixture also had medicinal purposes and was an important part of many traditional ceremonies.

Christopher Columbus was introduced to cacao and chocolate on his last journey to the Americas in 1502. He took some beans back to Spain to show the king and queen, however, they were viewed with apathy. Chocolate didn't truly arrive in Europe until 1585, when a shipment of beans came from Mexico to Spain. At this stage chocolate was still served as a drink but the Spaniards replaced the chilli with milk and sugar to sweeten the bitter taste. Cacao beans were in short supply and Spain guarded the secret of chocolate jealously. However, the luxury began to spread across the rest of Europe during the 17th century. Italy was next to appreciate chocolate and it finally arrived in England in 1650. Chocolate was only available to the wealthiest, who consumed it in fashionable 'chocolate houses', much like today's coffee shops. The first chocolate house opened in London in 1657.

Chocolate remained a beverage for almost two hundred more years. It wasn't until 1847 that Joseph Fry created the first solid chocolate bar for eating. Others followed soon after, with John Cadbury adding a chocolate bar to his range. The solid chocolate bar was based on cocoa butter, extracted from cacao beans in a method developed by the Dutch chocolate maker, Casparus van Houten in 1828. Chocolate was still dark at this point and there was not quite the selection that we have today. In 1875 Daniel Peter produced the first milk chocolate bar using powdered milk; he was assisted in his work by Henri Nestle, a name still affiliated with chocolate today.

The world continues to be obsessed with chocolate. We still love to eat chocolate, and Switzerland currently consumes the most at 10kg per person each year. Contemporary chefs not only produce chocolate based desserts and puddings but also combine sweet and savoury by adding chocolate to main course dishes, to stews, meat pies and even brussel sprouts. Modern society has also found other uses for chocolate. Chocolate face masks and chocolate massages are just some of the inventive ways in which chocolate has been used in health spas and beauty salons! It seems that chocolate, though now far removed from that known to the Aztecs, is as relished now as it was then.

Appendix H: Written precis scoring schedule

General rules

- Score 1 for each point (total of 20 points)
- Please note that spelling errors DO NOT affect scoring, except for the distinction between **cacao** and **cocoa** in **cacao beans** and **cocoa butter (e.g., cocoa beans = error).**
- Additionally, to gain a score of 1 each point must be included within the correct context.
- Slashes do signal optionality, apart from **question 4**, with 'currency/ medicinal purposes/ traditional ceremonies' where at least 2 out of the three items need to be mentioned, and apart from **question 20** where both items must be mentioned to get 1 point (both: masks and massages).
- There are no half points so both items must be reported e.g., Aztecs and Mayan. If they mentioned Aztecs only (or Mayan only) no points should be given.

Content Point	**Rules for point allocation**
1. Central and South America	Allocate one point only if both: *Central* and *South America* are listed. • Central can be replaced with: ***centre, middle*** *America* • *Central and South American**s*** (as if of the people) are acceptable. • ***Central and south Indian tribes/Latin America***- these are NOT correct. Spelling and capital letters have no impact on scoring here.
2. Aztecs and Mayans	Allocate one point if both: ***Aztecs and Mayans*** are mentioned, it does not have to be in one phrase such as: *Aztecs and Mayans prepared special drinks*. Names of both tribes can be separated within/ across sentences, e.g., ***Mayans*** *originally viewed cacoa beans as a gift from God and* ***Aztecs*** *associated them with fertility.* Spelling and capital letters do not influence scoring. I accepted examples such as ***Mayans, Mayas, Azteks, Aztec(s)*** etc.
3. Cacao beans	Allocate one point if the exact phrase is present, correctly spelled, in the right context. **NOTE**: this is the only example where spelling matters.
4. Currency/medicinal purposes/traditional ceremonies	Allocate one point if at least two out of the three concepts are listed. **Note:** *fertility* should not be interpreted as medicinal purposes. Synonyms are acceptable. For example: • ***currency*** = *coins*, money • ***medicinal purposes =*** *medicine/health purposes/healing properties*

	• ***traditional ceremonies =*** *ceremonies/ rituals/ rituals to the gods/ connection to the gods/ offerings/sacrifice object/sacrificial object/ sacred item /they've worshipped them/ religion, culture, tradition/ religious, cultural, traditional, spiritual, ritual (significance, value, importance, connotations, purposes, meaning, perspective, usage, symbol, event)*
5. Made from water and chilli spice/chilli	Allocate one point only if both items are present. I give point for: ***water and chilli***/ / ***water and chilli spice*** No point for: ***water and spice***, ***a drink from chillies*** *etc.*
6. Christopher Columbus/Columbus	No major issues with this one, frequently present, either the surname or both, the name and surname. Spelling and lack of capital letters do not affect scoring.
7. He took some beans back to Spain/back to show the King & Queen	They need to mention either: • *took some beans back to Spain OR,* • *took some beans to show the King and Queen.* **<u>This one may be a bit tricky:</u>** ***Take back*** can be replaced with *imported, presented, introduced*, e.g., *It was first exported to Spain in 1502 by <u>Columbus</u>*. ***Beans****:* must be mentioned, it cannot be replaced with *chocolate* e.g. *he took chocolate back to Spain* - is not good. OR the fact that these are beans must be inferred from the context. Beans can be replaced with ***cacao/ plant****.* ***Spain*** must be mentioned, it cannot be replaced with e.g. *Europe, continent* OR, the fact that this is Spain must be inferred from the context. ***King and queen*** can be replaced with synonyms: *royal family/ royals/ Spanish Royals/ monarchs of Spain/ royal house of Spain/ royalty.* If **Columbus** is not mentioned anywhere in the text and then something like: *they were taken to Spain, someone took them to Spain*, it is **not** OK**.**
8. King and Queen did not like it/viewed with apathy/apathetic/disregarded/not interested/did not care/indifferent	Any of those formulations would do, as they all describe the same concept ***King and Queen*** = *royal family/ royals/ Spanish Royals/ monarchs of Spain/ royal house of Spain/ royalty*

	King and Queen **cannot** be replaced with: *Spanish Monarchy* (a reference to a country rather than to the people), *King* (itself, and queen not mentioned).
9. Spaniards replaced chilli with milk and sugar/added milk and sugar	*added milk and sugar* is enough for a point, they do not need to mention what happened to chilli. So either: *replaced the chill with*.... OR *added milk and sugar* are acceptable answers. ***Sugar*** *can be replaced with* ***sweetener/ sweet.*** ***Spaniards*** can be replaced with ***Spanish people***; it cannot be replaced with ***Europeans*** (but if you can infer from the text hat Europeans actually mean Spaniards- this is OK.
10. Spain guarded secret/secret guarded/guarded jealously	Focus on *guarding* - not essential to mention *jealously*. Focus should be on the fact that they guarded it. ***jealously*** does not have to mentioned ***guarded*** can also be replaced with synonyms Again, infer from the context. e.g.: *Spaniards became protective over the secret of Chocolate,* this is OK.
11. Luxury/regarded as a luxury/wealthy/expensive /rich people	Either is acceptable. No bigger issues with this one. Acceptable synonyms: ***valuable, affluent, upper classes, high class people, elites/fortunate/priviledged people***
12. Spread to Italy and England/spread to Europe	Either are acceptable. ***Europe*** can be replaced by *continent* (if you can infer from the text that this is the European continent). ***Spread***, this word does not have to be mentioned but only suggested in the text, e.g. by expressions like the following: • *it became popular in Europe* • *Italy was the next country to be introduced to chocolate and the substance did not reach England until 1847* • *and soon it became popular across western Europe* • *… and it moved to Italy in the 17th century. After reaching Italy, it made it to England in 1650's.* • *It was a secret guarded for a long time until Italy discovered it then England.*
13. Joseph Fry/Fry	Either is good (surname or name and surname) No bigger issues with this one, make sure that the context is correct because people tend to confuse facts connected to this person.
14. John Cadbury/Cadbury	Either is good (surname or name and surname). As above (make sure that the context is OK).

15. Solid bar	The exact phrase must be present. This is usually referred to as: *solid chocolate bar*, which also is OK. Acceptable synonyms: *solid block, bar in solid form*
16. Powdered milk	***Milk powder***, very commonly used, this is acceptable. Make sure that the context is correct.
17. Henri Nestle	No major issue with this but both: name and the surname must be mentioned in order to get a point.
18. Savoury/main course/main dish	Optional, no issues with this one. ***savoury dish*** can be replaced with *saltier dish* ***main dish*** can be replaced with *main meal*
19. Today/modern society/now	Also acceptable: • *nowadays* • *in the modern age* • *in modern times/ in modern day* • *to this day* • *modern uses* • *currently/ in the current day,* • *modern chocolate consumption/ modern uses of chocolate* • *most recently* • *in these days* • *in the 21st century* *Today*- is also used in another context which does not count as a point, e.g., *coffee shops as we know today, the way we eat it today etc.*
20. Face masks/chocolate face masks/massages/ chocolate massages	Must mention <u>the masks</u> AND <u>massages</u> but can use either phrasing to refer to the two. ***face masks*** can be replaced with *facials*

References:

Abad, F., Quiroga, M. A., & Colom, R. (2016). Intelligence assessment. In Encyclopedia of applied psychology. Online reference database titled Neuroscience and biobehavioral psychology. Oxford, UK: Elsevier Ltd.

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, 15, 163–181. https://doi.org/10.1037/a0015719

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16, 270–301. https://doi.org/10.1177/1094428112470848

Alderson, J. C. (1993). The relationship between grammar and reading in an English for academic purposes test battery. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium*. Alexandria, VA: TESOL.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

Al-Malki, M. A. S. (2014). Testing the predictive validity of the IELTS test on Omani English candidates' professional competencies. *International Journal of Applied Linguistics and English Literature*, 3, 166–172. http://dx.doi.org/10.7575/aiac.ijalel.v.3n.5p.166

Al-Musawi, N. M., & Al-Ansari, S. H. (1999). Test of English as a Foreign Language and First Certificate of English tests as predictors of academic success for undergraduate students at the University of Bahrain. *System*, 27, 389–39. https://doi.org/10.1016/S0346-251X(99)00033-0

Alsager, R., & Milton, J. (2016). Investigating the relationship between vocabulary knowledge and academic success of Arabic undergraduate learners in Swansea university. *Language in Focus*, 2, 88–124. https://doi.org/10.1515/lifijsal-2016-0010

Amirjalili, F., & Jabbari, A. A. (2018). The impact of morphological instruction on morphological awareness and reading comprehension of EFL learners. *Cogent Education*, *5*, 1–30. https://doi.org/10.1080/2331186X.2018.1523975

Andrade, M. S. (2006). International students in English-speaking universities: Adjustment factors. *Journal of Research in International education*, 5, 131–154. https://doi.org/10.1177/1475240906065589

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using zoom videoconferencing for qualitative data collection: Perceptions and experiences of researchers and participants. *International Journal of Qualitative Methods*, 18, 1–8. https://doi.org/10.1177/1609406919874596

Atkins, W. B. & Baddeley, A. D. (1998). Working memory and distributed vocabulary learning. *Applied Psycholinguistics*, 19, 537–552. https://doi.org/10.1017/S0142716400010353

Ayers, J. B., & Peters, R. M. (1977). Predictive validity of the Test of English as a Foreign Language for Asian graduate students in engineering, chemistry, or mathematics. *Educational and Psychological Measurement*, 37, 460–463. https://doi.org/10.1177/001316447703700221

Ayers, J. B., & Quattlebaum, R. F. (1992). TOEFL performance and success in a master's program in engineering. *Educational and Psychological Measurement*, 52, 973–975. https://doi.org/10.1177/0013164492052004021

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4, 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. (1974). Working memory revised. *American Psychologist*, 19, 851–864.

Barnett, M. A. (1986). Syntactic and lexical/semantic skills in foreign language reading: importance and interaction. *Modern Language Journal*, 70, 343–9. https://doi.org/10.2307/326811

Barton, B., & Neville-Barton, P. (2003, February). Investigating the relationship between English language and mathematical learning. In: *Proceedings of the Third Conference of the European Society for Research in Mathematics Education, 28 February-3 March* (p. 1–10).

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 27, 101–118. https://doi.org/10.1177/0265532209340194

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, 16, 131–162. https://doi.org/10.1177/026553229901600202

Bellingham, L. (1995). The relationship of language proficiency to academic success for international students. *New Zealand Journal of Educational Studies*, 30, 229–232.

Berman, R., & Cheng, L. (2001). English academic language skills: Perceived difficulties by undergraduate and graduate students, and their academic achievement. *Canadian Journal of Applied Linguistics*, 4, 25–40. https://journals.lib.unb.ca/index.php/CJAL/article/view/19830

Bers, T. (1994). English proficiency, course patterns, and academic achievements of limited-English-proficient community college students. *Research in Higher Education*, 35, 209–234.

Brackenbury, W., J. (2019). Relationship between lecture capture usage and examination performance of Biology undergraduates. *York Scholarship of Teaching and Learning English*, 2, 11– 22. https://doi.org/10.1080/00219266.2019.1707258

Bridgeman, B., Cho, Y., & DiPietro, S. (2015). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing*, 3, 307–318. https://doi.org/10.1177/0265532215583066

Brisbois, J. E. (1995). Connections between first- and second-language reading. *Journal of Reading Behaviour*, 27, 565–84. https://doi.org/10.1080/10862969509547899

Brooks, G., & Adams, M. (2002). Spoken English proficiency and academic performance: Is there a relationship and if so, how do we teach? *Business Communication Quarterly* 67, 294–307.

Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2022). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, 43, 596–602. https://doi.org/10.1093/applin/amaa061

Brown, J. I., Fishco, V. V., & Hanna, G. (1993). *Nelson-Denny Reading Test (Forms G & H).* Chicago, IL: Riverside.

Bruno, R. M., & Walker, S. C. (1999). Comprehensive test of phonological processing (CTOPP). *Diagnostique*, 24, 69–82. https://doi.org/10.1177/153450849902401-408

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual differences*, 29, 1057–1068. https://doi.org/10.1016/S0191-8869(99)00253-6

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin,* 40, 153−193. https://doi.org/10.1037/h0059973

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition*, 28, 143−153.

Chervenkova, V. (2021). International Facts and Figures 2021. Universities UK. www.universitiesuk.ac.uk

Chia, C. K., Ghavifekr, S., & Razak, A. Z. A. (2020). Online interview tools for qualitative data collection during Covid-19 pandemic; Riview of web conferencing platforms' functionality. *Malaysian Journal of Qualitative Research*, 7, 95–106.

Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S. J. (2019). The matrix reasoning item bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society Open Science*, 6, 1–13. 190232. https://doi.org/10.1098/rsos.190232

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing*, *29*, 421–442. https://doi.org/10.1177/0265532211430368

Chun, D. M., & Payne, J. S. (2004). What makes students click: Working memory and look-up behavior. *System*, 32, 481–503. https://doi.org/10.1016/j.system.2004.09.008

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge Academic.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research Methods in Education*. Routledge.

Coleman, J. A. (2004). Study abroad. In: M. Byram (Ed.), *Routledge Encyclopaedia of Language Teaching and Learning* (pp. 582–584). London: Routledge. https://doi.org/10.4324/9780203219300

Coley, M. (1999). The English language entry requirements of Australian Universities for students of non-English speaking background. *Higher Education Research & Development*, 18, 7–17. https://doi.org/10.1080/0729436990180102

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *IELTS Research Reports*, 1, 72–115.

Crawford, I., Wang, Z. (2014). Why are first-year accounting studies inclusive? *Accounting & Finance*, 54, 419–439. https://doi.org/10.1111/acfi.12007

Crawford, I., Wang, Z. (2015). The impact of individual factors on the academic attainment of Chinese and UK students in higher education. *Studies in Higher Education*, 40, 902–920. https://doi.org/10.1080/03075079.2013.851182

Crystal, D. (1997). *The Cambridge Encyclopaedia of Language*. Cambridge: Cambridge University Press.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A Reassessment1. *Applied linguistics*, 2, 132−149.

Dąbrowska, E. (2018). Experience, aptitude and individual differences in native language ultimate attainment. *Cognition*, 178, 222−235. https://doi.org/10.1016/j.cognition.2018.05.018

Daller, M., Kuiken, F., Trenkic, D., & Vedder, I. (2021). Linguistic predictors of academic achievement amongst international students and home students in higher education: introduction. *International Journal of Bilingual Education and Bilingualism*, 24, 1453–1457.

Daller, M., Müller, A., & Wang-Taylor, Y. (2021). The C-test as predictor of the academic success of international students. *International Journal of Bilingual Education and Bilingualism*, 24, 1502–1511.

Daller, M. H., Phelan, D. (2013). Predicting international student study success. *Applied Linguistics Review*, 4, 173–193. https://doi.org/10.1515/applirev-2013-0008

Daller, M. H., Xue, H. (2009). Vocabulary knowledge and academic success: A study of Chinese students in UK higher education. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, J. Treffers- Daller (Eds.), *Vocabulary Studies in First and Second Language Acquisition* (pp. 179−193). Palgrave Macmillan, London. https://doi.org/10.1057/9780230242258_11

Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450−466. https://doi.org/10.1016/S0022-5371(80)90312-6

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language*, 25, 1−18. https://doi.org/10.1016/0749-596X(86)90018-5

Daniels, N., Gillen, P., Casson, K., & Wilson, I. (2019). STEER: Factors to consider when designing online focus groups using audio-visual technology in health research. *International Journal of Qualitative Methods*, 18, 1–11. https://doi.org/10.1177/1609406919885786

De Vita, G. (2002). Cultural equivalence in the assessment of home and international business management students: A UK exploratory study. *Studies in Higher Education*, 27, 221–231. https://doi.org/10.1080/03075070220120038

Devos, N. J. (2019). Comparing first-term students' English language proficiency at a Canadian polytechnic institute. *BC TEAL Journal*, 4, 53–83. https://doi.org/10.14288/bctj.v4i1.335

Di Fabio, A., & Palazzeschi, L. (2009). An in-depth look at scholastic success: Fluid intelligence, personality traits or emotional intelligence? *Personality and Individual Differences*, *46*, 581–585. https://doi.org/10.1016/j.paid.2008.12.012

Dillon, A. (1992). Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics*, 35, 1297–1326.

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, 4, 500–503. https://doi.org/10.1037/0882-7974.4.4.500

Edwards, V., Ran, A., & Li, D. (2007). Uneven playing field or falling standards? Chinese students' competence in English. *Race Ethnicity and Education*, 10, 387–400. https://doi.org/10.1080/13613320701658431

Elder, C. (1993). Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Language Testing*, 2, 68–85.

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly,* 5, 173–194. https://doi.org/10.1080/15434300802229334

Elder, C., Bright, C., & Bennett, S. (2007). The role of language proficiency in academic success: Perspectives from a New Zealand university. *Melbourne Papers in Language Testing*, 12, 24–28.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*, 253–272. https://doi.org/10.1177/0265532212459028

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., Van Daal, V. H., Polyzoe, N., ... & Petalas, M. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly*, 39, 438–468.

Evans, S., & Morrison, B. (2011). Meeting the challenges of English-medium higher education: The first-year experience in Hong Kong. *English for Specific Purposes*, 30, 198–208. https://doi.org/10.1016/j.esp.2011.01.001

Farsides, T., & Woodfield, R. (2003). Individual differences and undergraduate academic success: The roles of personality, intelligence, and application. *Personality and Individual differences*, 34, 1225–1243. https://doi.org/10.1016/S0191-8869(02)00111-3

Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal*, 3, 70–85. https://doi.org/10.1177/1475240909358529

Ferguson, G., & White, E. (1998). A small-scale study of predictive validity. *Melbourne Papers in Language Testing*, *7*, 15–63.

Field, A. (2009). Discovering Statistics Using SPSS. (3rd edition). London: Sage.

Field, A., & Hole, G. (2002). How to Design and Report Experiments. London: Sage.

Fiocco, M. (1992). English proficiency levels of students from a non-English speaking background: a study of IELTS as an indicator of tertiary success. Unpublished research report. Perth: Curtin University of Technology.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of memory and language*, 41, 78–104. https://doi.org/10.1006/jmla.1999.2638

Fox, J. (2005). Rethinking second language admission requirements: Problems with language-residency criteria and the need for language assessment and support. *Language Assessment Quarterly: An International Journal*, 2, 85–115. https://doi.org/10.1207/s15434311laq0202_1

Fuertes, J. N., Gottdiener, W. H., Martin, H., Gilbert, T. C., & Giles, H. (2012). A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology*, 42, 120–133. https://doi.org/10.1002/ejsp.862.

Furnham, A., & Chamorro-Premuzic, T. (2004). Personality and intelligence as predictors of statistics examination grades. *Personality and individual differences*, 37, 943–955. https://doi.org/10.1016/j.paid.2003.10.016

Gardner, H., & Moran, S. (2006). The science of multiple intelligences theory: A response to Lynn Waterhouse. *Educational Psychologist*, *41*, 227–232. https://doi.org/10.1207/s15326985ep4104_2

Geelhoed, R. J., Abe, J., & Talbot, D., M. (2003). A qualitative investigation of US students' experiences in an international peer program. *Journal of College Student Development*, 44, 5–17. https://doi.org/10.1353/csd.2003.0004

Georgiou, G. K., & Das, J. P. (2018). Direct and indirect effects of executive function on reading comprehension in young adults. *Journal of Research in Reading*, 41, 243–258. https://doi.org/10.1111/1467-9817.12091

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing*, 35, 271–295. https://doi.org/10.1177/0265532217704010

Gluszek, A., & Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review,* 14, 214–237. https://doi.org/10.1177/1088868309359288.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? Applied Linguistics, 11, 341–363. https://doi.org/10.1093/applin/11.4.341

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly*, 21, 505–521. https://doi.org/10.2307/3586500

Grayson, J. P. (1997). Academic achievement of first-generation students in a Canadian university. *Research in higher Education*, *38*, 659–676. https://doi.org/10.1023/A:1024955719648

Grayson, J. P. (2008a). The experiences and outcomes of domestic and international students at four Canadian universities. *Higher Education Research & Development*, 27, 215–230. https://doi.org/10.1080/07294360802183788

Grayson, J. P. (2008b). Sense of coherence and academic achievement of domestic and international students: A comparative analysis. *Higher education*, *56*, 473–492. https://doi.org/10.1007/s10734-007-9106-0

Grayson, J., & Stowe, S. (2005). Language problems of international and domestic ESL students at UBC, York, McGill and Dalhousie, and Academic Achievement. In: Annual conference of the Canadian Society for the Study of Education, London, Ontario, Canada.

Gu, Q., & Maley, A. (2008). Changing places: A study of Chinese students in the UK. *Language and Intercultural Communication*, 8, 224–245. https://doi.org/10.1080/14708470802303025

Gue, L. R., & Holdaway, E. A. (1973). English proficiency tests as predictors of success in graduate studies in education. *Language Learning*, *23*, 89–103. https://doi.org/10.1111/j.1467-1770.1973.tb00099.x

Gyllstad, H., McLean, S., & Stewart, J. (2021). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, *38*, 558–579.

Hakuta, K., Butler, Y. G., & Witt, D. (2000). How long does it take English learners to attain proficiency? The University of California Linguistic Minority Research Institute Policy Report.

Harrington, M., & Roche, T. (2014). Identifying academically at-risk students in an English-as-a-Lingua-Franca university setting. *Journal of English for Academic Purposes*, 15, 37–47. https://doi.org/10.1016/j.jeap.2014.05.003

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, 14, 25–38.

Harris, D. (1940). Factors affecting college grades. A review of the literature. *Psychological Bulletin*, 37, 125–166. https://doi.org/10.1037/h0055365

Harrison, G. L., Goegan, L. D., Jalbert, R., McManus, K., Sinclair, K., & Spurling, J. (2016). Predictors of spelling and writing skills in first-and second-language learners. *Reading and Writing*, 29, 69–89. https://doi.org/10.1007/s11145-015-9580-1

Hartnett, N., Romcke, J., & Yap, C. (2004) Student performance in tertiary-level accounting: An international student focus. *Accounting and Finance*, 44, 163–85. https://doi.org/10.1111/j.1467-629X.2004.00104.x

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, 177, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Hashimoto, B. J. (2021). Is frequency enough? The frequency model in vocabulary size testing. *Language Assessment Quarterly*, 18, 171–187. https://doi.org/10.1080/15434303.2020.1860058

Havas, V., Taylor, J. S. H., Vaquero, L., de Diego-Balaguer, R., Rodriguez-Fornells, A., & Davis, M. H. (2018). Semantic and phonological schema influence spoken word learning and overnight consolidation. *Quarterly Journal of Experimental Psychology,* 71, 1469−1481. https://doi.org/10.1080/17470218.2017.1329325

Haynes, M., & Carr, T. H. (1990). Writing system background and second language reading: a component skills analysis of English reading by native speaker-readers of Chinese. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: component skills approaches*. San Diego, CA: Academic Press.

He, Y., & Banham, H. (2009). International student academic performance: Some statistical evidence and its implications. *American Journal of Business Education (AJBE)*, 2, 89−100. https://doi.org/10.19030/ajbe.v2i5.4073

HEPI. (2021). Summary report for the Higher Education Policy Institute and Universities UK International. https://www.hepi.ac.uk/

HESA. (2023). Higher education statistics for the United Kingdom 2020–21. Higher Education Statistics Agency. https://www.hesa.ac.uk/

Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *International English Language Testing System (IELTS) Research Reports 1999,* 2, 62–73.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and writing*, 2, 127–160. https://doi.org/10.1007/BF00401799

Horrocks, G. (1987). *Generative Grammar*. London: Longman.

Hsueh-Chao, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13, 403–430.

Hu, R., & Trenkic, D. (2021). The effects of coaching and repeated test-taking on Chinese candidates' IELTS scores, their English proficiency, and subsequent academic

achievement. *International Journal of Bilingual Education and Bilingualism*, 24, 1486–1501. https://doi.org/10.1080/13670050.2019.1691498

Hwang, K. Y., & Dizney, H. F. (1970). Predictive validity of the test of English as a foreign language for Chinese graduate students at an American university. *Educational and Psychological Measurement*, 30, 475-477.

Iannelli, C., & Huang, J. (2014). Trends in participation and attainment of Chinese students in UK higher education. *Studies in Higher Education*, 39, 805–822. https://doi.org/10.1080/03075079.2012.754863

ICEF Monitor**.** (2022). Germany's foreign enrolment reaches record high with increase of 8% in 2021/22. https://monitor.icef.com/2022/10/germanys-foreign-enrolment-reaches-record-high-with-increase-of-8-in-2021-22/

IELTS Partners. (2017). IELTS for study. https://www.ielts.org/what-is-ielts/ielts-for-study

IELTS. (2017). Guide for educational institutions, governments, professional bodies, and commercial organisations. www.ielts.org

Ingram, D., & Bayliss, A. (2007). IELTS as a predictor of academic language performance, Part 1. *IELTS Research Reports*.

International Baccalaureate. (2022). Benchmarking diploma programme language courses to the CEFR. www.ibo.org

International Education Strategy. (2021). Supporting recovery, driving growth. https://www.gov.uk

Isonio, S. (1994). Retention and success rates by course category, year, and selected student characteristics at Golden West College. Huntington Beach, CA: Golden West College.

Jenkins, J., & Wingate, U. (2015). Staff and students' perceptions of English language policies and practices in 'international' universities: a case study from the UK. *Higher Education Review*, 47, 47–73.

Jiang, X., Sawaki, Y., & Sabatini, J. (2012). Word reading efficiency, text reading fluency, and reading comprehension among Chinese learners of English. *Reading Psychology*, 33, 323–349. https://doi.org/10.1080/02702711.2010.526051

Jin, L., & Cortazzi, M. (1996). This way is very different from Chinese ways: EAP needs and academic culture, *Review of ELT*, 6, 205–217.

Jin, L., & Cortazzi, M. (2006). Changing practices in Chinese cultures of learning. *Language, Culture and Curriculum*, 19, 5–20. https://doi.org/10.1080/07908310608668751

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive psychology*, 21, 60–99. https://doi.org/10.1016/0010-0285(89)90003-0

Johnson, P. (1988). English language proficiency and academic performance of undergraduate international students. *TESOL Quarterly*, 22, 164–168.

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports*, 3, 86–108.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, 57, 1–44. https://10.1111/0023-8333.101997010-i1

Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, 123, 138–149. https://doi.org/10.1016/j.compedu.2018.05.005

Krausz, J., Schiff, A., Schiff, J., & Hise, J. V. (2005). The impact of TOEFL scores on placement and performance of international students in the initial graduate accounting class. *Accounting Education*, 14, 103–111. https://doi.org/10.1080/0963928042000256671

Laidra, K., Pullmann, H., & Allik, J. (2007). Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and Individual Differences*, 42, 441–451. https://doi.org/10.1016/j.paid.2006.08.001

Lam, D. M. K., Green, A., Murray, N., & Gayton, A. (2021.) How are IELTS scores set and used for university admissions selection: A cross-institutional case study. IELTS Research Reports Online Series, 3. British Council, Cambridge Assessment English and IDP: IELTS Australia. https://www.ielts.org/-/media/research-reports/lam-et-al-report-layout-april-2021.ashx

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Bejonit. *Vocabulary and Applied Linguistics,* (p. 126–132). Palgrave Macmillan, London.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, 19, 255–271. https://doi.org/10.1093/applin/19.2.255

Lesser, J. M. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning*, 57, 229–270. https://doi.org/10.1111/j.1467-9922.2007.00408.x

Li, C. N., & Thompson, S. A. (2003). *The world's major languages*. Routledge.

Li, G., Chen, W., & Duanmu, J. L. (2010). Determinants of international students' academic performance: A comparison between Chinese and other international students. *Journal of Studies in International Education*, 14, 389–405. https://doi.org/10.1177/1028315309331490

Light, R. L., Xu, M., & Mossop, J. (1987). English proficiency and academic performance of international students. *TESOL Quarterly*, *21*, 251–261. https://doi.org/10.2307/3586734

Lightbown, P., & Spada, N. (2013). *How languages are learned* (4th ed.). Oxford, UK: Oxford University Press.

Lobe, B., Morgan, D., & Hoffman, K. A. (2020). Qualitative data collection in an era of social distancing. *International Journal of Qualitative Methods*, 19, 1–8. https://doi.org/10.1177/1609406920937875

Lounsbury, J. W., Sundstrom, E. D., Loveland, J. L., & Gibson, L. W. (2003). Intelligence, ''Big Five" personality traits and Work Drive as predictors of course grade. *Personality and Individual Differences*, 35, 1231–1239. https://doi.org/10.1016/S0191-8869(02)00330-6

Luo, J., & Jamieson-Drake, D. (2013). Examining the educational benefits of interacting with international students. *Journal of International Students*, 3, 85–101.

Ma, Q., & Kelly, P. (2009). Overcoming hurdles to Chinese students' learning of English lexis. *Changing English*, 16, 405–412. https://doi.org/10.1080/13586840903391997

Makepeace, E. & Baxter, A. (1990). Overseas students and examination failure: a national study. *Journal of International Education*, 1, 36–48.

Manning, A. (2018). Impact of international students in the UK. Digital Education Resource Archive. https://dera.ioe.ac.uk/32233/

Marinus, E., Kohnen, S., & McArthur, G. (2013). Australian comparison data for the Test of Word Reading Efficiency (TOWRE). *Australian Journal of Learning Difficulties*, 18, 199–212. https://doi.org/10.1080/19404158.2013.852981

Masrai, A., & Milton, J. (2017). Recognition vocabulary knowledge and intelligence as predictors of academic achievement in EFL context. *TESOL International Journal*, 12, 128–142.

Masterson, J., & Hayes, M. (2004). Development and data for UK versions of an author and title recognition test for adults. *Journal of Research in Reading*, 30, 212–219.

Matthews, K. L., Baird, M., & Duchesne, G. (2018). Using online meeting software to facilitate geographically dispersed focus groups for health workforce research. *Qualitative Health Research*, 28, 1621–1628. https://doi.org/10.1177/1049732318782167

Mayor, B. M. (2006). Dialogic and hortatory features in the writing of Chinese candidates for the IELTS test. *Language, Culture and Curriculum*, 19, 104–121. https://doi.org/10.1080/07908310608668757

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23, 412–433. https://doi.org/10.1037/met0000144

Mehrpour, S., Razmjoo, S. A., & Kian, P. (2011). The relationship between depth and breadth of vocabulary knowledge and reading comprehension among Iranian EFL learners. *Journal of English Language Teaching and Learning*, 53, 97–127.

Melby-Lervåg, M., & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, 140, 409–433. https://doi.org/10.1037/a0033890

Mellors-Bourne, R., Humphrey, C., Kemp, N., & Woodfield, S. (2013). The wider benefits of international higher education in the UK. Department for Business Innovation & Skills.

Mendelsohn, D. (2002). The lecture buddy project: An experiment in EAP listening comprehension. *TESL Canada Journal*, 20, 64–73. https://doi.org/10.18806/tesl.v20i1.939

Milton, J. (2009). Measuring second language vocabulary acquisition. Bristol, UK: Multilingual Matters.

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL-International Journal of Applied Linguistics*, 107, 17– 34.

Mitchell, J. J. (2001). Comprehensive test of phonological processing. *Assessment for Effective Intervention*, 26, 57-63.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, 28, 291–304.

Montgomery, C. (2009). A decade of internationalisation: has it influenced students' views of cross-cultural group work at university? *Journal of studies in international education*, 13, 256–270. https://doi.org/10.1177/1028315308329790

Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, 4, 43–66. https://doi.org/10.1016/j.jeap.2004.02.001

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of teaching English as a Second Language trainees. *System*, 32, 75–87. https://doi.org/10.1016/j.system.2003.05.001

Morrison, J., Merrick, B., Higgs, S., & Le Métais, J. (2005). Researching the performance of international students in the UK. *Studies in Higher Education*, 30, 327–337. https://doi.org/10.1080/03075070500095762

Müller, A., & Daller, M. (2019). Predicting international students' clinical and academic grades using two language tests (IELTS and C-test): A correlational research study. *Nurse Education Today*, 72, 6–11.

Murray, N. (2010). Considerations in the post-enrolment assessment of English language proficiency: Reflections from the Australian context. *Language Assessment Quarterly*, 7, 343–358. https://doi.org/10.1080/15434303.2010.484516

Murray, N. (2018). University gatekeeping tests: What are they really testing and what are the implications for EAP provision? *J-Stage Journal*, 62, 15–27. https://doi.org/10.32234/jacetjournal.62.0_15

Na, L. & Nation, I., S., P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal,* 16, 33–42. https://doi.org/10.1177/003368828501600103

Nasirudeen, A. M. A., & Xiao, S. (2020). English language skills and academic performance: A comparison between Asian international and domestic nursing students in Singapore. *International Journal of Nursing*, 7, 30–38.

Nathan, R. G., & Stanovich, K. E. (1991). The causes and consequences of differences in reading fluency. *Theory into Practice*, 30, 176–184.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review,* 63, 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P., Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31, 9–13.

Nation, P., & Newton, J. (1997). Teaching vocabulary. In J. Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition*. Cambridge, UK: Cambridge University Press.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), Vocabulary: Description, acquisition and pedagogy, p. 6–19. Cambridge: Cambridge University Press.

Neal, M. (1998). The predictive validity of the GRE and TOEFL exams with GGPA as the criterion of graduate success for international graduate students in science and engineering. Retrieved from EBSCOhost.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J. & Urbina, S. (1996). Intelligence. *American Psychologist*, 51, 77–101. https://doi.org/10.1037/0003-066X.51.2.77

Nguyen, L., T., C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, 42, 86–99. https://doi.org/10.1177/0033688210390264

Nielsen, J. (1997). Why users scan instead of reading. Nielsen Norman Group. www.nngroup.com

Nielsen, J. (2010). iPad and Kindle reading speeds. Nielsen Norman Group. www.nngroup.com

O'Reilly, D., & Marsden, E. (2021). Eliciting and measuring L2 metaphoric competence: Three decades on from Low (1988). *Applied Linguistics*, 42, 24–59. https://doi.org/10.1093/applin/amz066

O'Connell, R. M., & Resuli, N. (2020). Academic challenges for Chinese transfer students in engineering. *Journal of International Students*, 10, 466–482. https://doi.org/10.32674/jis.v10i2.674

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524537

Olive, V. (2017). How much is too much? Cross-subsidies from teaching to research in British universities. HEPI report 100.

Oliver, R., Vanderford, S., & Grote, E. (2012). Evidence of English language proficiency and academic achievement of non-English-speaking background students. *Higher Education Research & Development*, 31, 541–555. https://doi.org/10.1080/07294360.2011.653958

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, 25, 26–45. https://doi.org/10.1017/S0267190505000024

Osaka, M., Osaka, N. & Groner, R. (1993). Language-independent working memory: Evidence from German and French reading Span tests. *Bulletin of the Psychometric Society*, 31, 117–118. https://doi.org/10.3758/BF03334156

Osaka, M., Osaka, N. (1992). Language independent working memory as measured by Japanese and English reading span tests. *Bulletin of Psychonomic Society*, 30, 287–289. https://doi.org/10.3758/BF03330466

Pasquarella, A., Chen, X., Gottardo, A., & Geva, E. (2015). Cross-language transfer of word reading accuracy and word reading fluency in Spanish-English and Chinese-English bilinguals: Script-universal and script-specific processes. *Journal of Educational Psychology*, 107, 96–110. https://doi.org/10.1037/a0036966

Pasquarella, A., Gottardo, A., & Grant, A. (2012). Comparing factors related to reading comprehension in adolescents who speak English as a first (L1) or second (L2) language. *Scientific Studies of Reading*, 16, 475–503. https://doi.org/10.1080/10888438.2011.593066

Patkowski, M., Fox, L., & Smodlaka, I. (1997). Grades of ESL and non-ESL students in selected courses in ten CUNY colleges. *College ESL*, 7, 1–13.

Paton, M. J. (2007). Why international students are at greater risk of failure. *International Journal of Diversity in Organizations, Communities & Nations*, 6, 101–112.

Pearson, W. S. (2021). The predictive validity of the Academic IELTS test: A methodological synthesis. *International Journal of Applied Linguistics*, 172, 85–120.

Pereltsvaig, A. (2012). *Languages of the world*. Cambridge: Cambridge University Press.

Peverly, S. T. (2006). The importance of handwriting speed in adult writing. *Developmental Neuropsychology*, 29, 197–216. https://doi.org/10.1207/s15326942dn2901_10

Plonsky, L. & Derrick, D. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, 100, 538–53. https://doi.org/10.1111/modl.12335

Podesva, R. J., & Sharma, D. (2014). *Research methods in linguistics*. Cambridge: Cambridge University Press.

Pringprom, P., & Obchuae, B. (2011). Relationship between vocabulary size and reading comprehension. *FLLT Proceedings*, 1, 182–191.

Purpura, J. E. (2004). *Assessing grammar.* Cambridge, UK: Cambridge University Press.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, 56, 282–308. https://doi.org/10.3138/cmlr.56.2.282

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language learning*, 52, 513–536. https://doi.org/10.1111/1467-9922.00193

Rankin, M., Silvester, M., Vallely, M., & Wyatt, A. (2003). An analysis of the implications of diversity for students' first level accounting performance. *Accounting & Finance*, 43, 365–393. https://doi.org/10.1111/j.1467-629x.2003.00096.x

Rashidi, N., & Khosravi, N. (2010). Assessing the role of depth and breadth of vocabulary knowledge in reading comprehension of Iranian EFL learners. *Journal of Pan-Pacific Association of Applied Linguistics*, 14, 81–108.

Read, J. (2008). Identifying academic language needs through diagnostic assessment. *Journal of English for Academic Purposes*, 7, 180–190. https://doi.org/10.1016/j.jeap.2008.02.001

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, 10, 77–89.

Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting. *Language Testing in Asia*, *3*, 1–13. https://doi.org/10.1186/2229-0443-3-12

Ross, C., & Ma, J. H. S. (2009). *Modern Mandarin Chinese grammar: A practical guide*. Routledge.

Ruspini, E. (2002). *Introduction to longitudinal research*. Psychology Press.

Sabourin, L., Leclerc, J. C., Lapierre, M., Burkholder, M., & Brien, C. (2016). The language background questionnaire in L2 research: Teasing apart the variables. In *Annual Meeting of the Canadian Linguistics Association, Calgary, Canada*.

Salthouse, T. A., Pink J., E., & Tucker-Drob, E., M. (2008). *Contextual analysis of fluid intelligence. Intelligence*, 36, 464–486. https://doi.org/10.1016/j.intell.2007.10.003

Schmidtke, D., & Moro, A. L. (2021). Determinants of word-reading development in English learner university students: A longitudinal eye movement study. *Reading Research Quarterly*, 56, 819–854. https://doi.org/10.1002/rrq.362

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, 64, 913–951. https://doi.org/10.1111/lang.12077

Senyshyn, R. M., Warford, M. K., & Zhan, J. (2000). Issues of adjustment to higher education: International students' perspectives. *International Education*, 30, 1–17.

Shahnazari-Dorcheh, M., & Adams, R. (2014). The relationship between working memory and L2 reading comprehension. *Applied Research on English Language*, 3, 19–34. https://doi.org/10.22108/are.2014.15492

Sharon, A. T. (1972). English proficiency, verbal aptitude, and foreign student success in American graduate schools. *Educational And Psychological Measurement*, 32, 425–431.

Shaw, P., & McMillion, A. (2018). Reading Comprehension in Advanced L2 Readers. In: K. Hyltenstam, I. Bartning & L. Fant (Eds.), *High-Level Language Proficiency in Second Language and Multilingual Contexts*. Cambridge University Press.

Shimamoto, T. (2000). An analysis of receptive vocabulary knowledge: Depth versus breadth. *JABAET Journal*, 4, 69–80.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, 24, 99–128. https://doi.org/10.1177/0265532207071513

Smrtnik Vitulić, H., & Prosen, S. (2012). Personality and cognitive abilities as predictors of university students' academic achievement. *Društvena istraživanja: časopis za opća društvena pitanja*, 21, 715–732. https://doi.org/10.5559/di.21.3.06

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36, 139–152. https://doi.org/10.1080/09571730802389975

Sternberg, R. J., & Kaufman, J. C. (1998). Human abilities. *Annual review of psychology*, 49, 479–502.

Storch, N., & Hill, K. (2008). What happens to international students' English after one semester at university? *Australian Review of Applied Linguistics*, 31, 4–17. https://doi.org/10.2104/aral0804

Sulistyo, G. H. (2009). TOEFL in a brief historical overview from PBT to IBT. *Bahasa Dan Seni*, 37, 116–127.

Szabo, C. Z., Stickler, U., & Adinolfi, L. (2021). Predicting the academic achievement of multilingual students of English through vocabulary testing. *International Journal of Bilingual Education and Bilingualism*, 24, 1531–1542. https://doi.org/10.1080/13670050.2020.1814196

Tarar, J. M., Meisinger, E. B., & Dickens, R. H. (2015). Test Review: Test of Word Reading Efficiency–Second Edition (TOWRE-2) by Torgesen, JK, Wagner, RK, & Rashotte, CA. *Canadian Journal of School Psychology*, 30, 320–326. https://doi.org/10.1177/0829573515594334

Thurlow, R., & van den Broek, P. (1997). Automaticity and inference generation during reading comprehension. *Reading and Writing Quarterly*, 13, 165–184. https://doi.org/10.1080/1057356970130205

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). Test of Word Reading Efficiency–Second Edition (TOWRE-2). Austin, TX: Pro-Ed.

Trenkic, D. & Warmington, M. (2019). Language and literacy skills of home and international university students: How different are they, and does it matter*? Bilingualism: Language and Cognition.* 22, 349–365. https://doi.org/10.1017/S136672891700075X

Trice, A. G. (2003). Faculty perceptions of graduate international students: The benefits and challenges. *Journal of Studies in International Education*, 7, 379–403. https://doi.org/10.1177/1028315303257120

UNESCO. (2019). Global Flow of Tertiary-Level Students. http://uis.unesco.org/en/uis-student-flow

UNESCO (2022). Higher Education global data report. Available online: https://bangkok.unesco.org/content/unesco-higher-education-global-data-report

van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first-and second-Language reading comprehension: A componential analysis. *Journal of educational psychology*, 96, 19–30. https://doi.org/10.1037/0022-0663.96.1.19

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22, 217–234. https://10.1017/S0142716401002041

Vinke, A. A., & Jochems, W. M. G. (1993). English proficiency and academic success in international postgraduate education, *Higher Education*, 26, 275–285. https://doi.org/10.1007/BF01383487

Volet, S. E., & Renshaw, P. D. (1996). Chinese students at an Australian university: Adaptability and continuity. In: D. A. Watkins & J. B. Biggs (Eds.), The Chinese learner: Cultural, psychological, and contextual influences (p. 205–220). Hong Kong University Press.

Wächter, B., & Maiworm, F. (2014). English-taught Programmes in European Higher Education: The State of Play in 2014. ACA Papers on International Cooperation in Education. Bonn: Lemmens.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (1999). Comprehensive test of phonological processing: CTOPP. Austin, TX: Pro-ed.

Warmington, M., Stothard, S. E., & Snowling, M. J. (2013). Assessing dyslexia in higher education: The York Adult Assessment Battery - Revised. *Journal of Research in Special Educational Needs*, 13, 48–56. https://doi.org/10.1111/j.1471-3802.2012.01264.x

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology*, 49, 51–79. https://doi.org/10.1080/713755607

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, 41, 85–120. https://doi.org/10.1177/00224669070410020401

Wechsler, D. (2011). Wechsler abbreviated scale of intelligence (2nd edition). San Antonio, TX: Pearson.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Woodrow, L. (2006). Academic success of international postgraduate education students and the role of English proficiency. *University of Sydney papers in TESOL*, 1, 51–70.

Yen, D., & Kuzma, J. (2009). Higher IELTS score, higher academic performance? The validity of IELTS in predicting the academic performance of Chinese students. *Worcester Journal of Learning and Teaching*, 3, 1–7.

Yixin, W., & Daller, M. (2014). Predicting Chinese students' academic achievement in the UK. Learning, Working and Communicating in a Global Context, 217, 217–227.

Yul, G., & Hoffman, P. (1990). Predicting success for international teaching assistant in a U.S. University. *TESOL Quarterly*, 24, 227–243.

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'anna, C. A., & Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior*, 27, 201–212. https://doi.org/10.1080/10862969509547878

Zeng, J. (1996). When east meets west: Mainland Chinese students and scholars in UK higher education institutions. *Journal of International Education*, 9, 9–15.

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 26, 696–725. https://doi.org/10.1177/1362168820913998

Zhao, J. C., & Mawhinney, T. (2015). Comparison of native Chinese-speaking and native English-speaking engineering students' information literacy challenges. *The Journal of Academic Librarianship*, 41, 712–724. https://doi.org/10.1016/j.acalib.2015.09.010