

Exploring the Predictors of Academic Vocabulary Proficiency among First-Year University Students

Memory Ponela (Student number: 1489805)

Supervisors:

Dr Ilse Fouché

Dr Laura Drennan



A report submitted in partial fulfilment of the requirements for the degree of

Master of Arts in the field of e-Science in the School of Social Sciences

University of the Witwatersrand, Johannesburg

14 July 2025

Abstract

Reading is an integral part of learning and supports academic success. Reports on the reading crisis in South Africa raise concerns about its impact on students' academic achievement. Integral to reading comprehension is vocabulary knowledge. This study explores the associations between academic vocabulary proficiency and various predictors among first-year university students. The study employs a three-stage analysis, which includes univariate analysis, bivariate analysis, and multiple linear regression analysis, to: (1) identify the predictors related to academic vocabulary proficiency, (2) explore the associations between academic vocabulary proficiency and each selected predictor, and (3) examine the extent of the relationships between academic vocabulary proficiency and the selected predictors. The findings reveal statistically significant and nuanced relationships between University Word List (UWL) scores and factors such as the reading strategies taught (or not taught) in high school, as well as proficiency at the 2000-, 3000-, and 5000-word levels. The study adds to the insights on academic reading and provides valuable insights for future research into the predictors of academic vocabulary proficiency and academic reading more broadly.

Declaration

I, Memory Ponela, declare that this report is my own, unaided work. It is being submitted for the degree of Master of Arts in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

A handwritten signature in black ink, appearing to read 'm. Ponela', written over a horizontal line.

Memory Ponela

14 July 2025

Acknowledgements

I would like to sincerely thank God for providing me with guidance and strength throughout my academic journey. I am deeply indebted to my supervisors Dr Ilse Fouche and Dr Laura Drennan. Thank you both so much for giving me an opportunity and providing me with this invaluable knowledge and insights. I would also like to extend to my appreciation to Dr Petra Gaylard; your guidance on the statistics section of this study is greatly appreciated.

To my family, I would like to extend my deepest appreciation for their unwavering support and understanding during this time. Your encouragement and belief in my abilities were a great source of motivation, giving me the strength to persevere through challenges and stay focused on my goals. To my friends, peers, and colleagues, I greatly appreciate your support, insightful conversations, and words of encouragement. Your willingness to listen, share ideas, and offer guidance has made a meaningful difference in this journey. I am truly grateful for the sense of community and belonging you have provided.

The e-Science programme the DST-CSIR National e-Science Postgraduate Teaching and Training Platform (NEPTTP) at the University of the Witwatersrand played a crucial role in enabling this research. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NEPTTP. Thank you to the e-Science lecturers and co-ordinators, Prof. Rod Alence, Dr Celso Monjane and Ms Casey Sparkes, for all the support and guidance.

Table of Contents

Chapter 1: Introduction - Background and Context	1
1.1 Introduction	1
1.2 Background and context	1
1.3 Statement of problem	5
1.4 Research question	6
1.5 Research objectives	6
1.6 Significance of study	6
1.7 Chapter outline	7
1.8 Chapter summary	7
Chapter 2: Literature Review and Conceptual Framework	8
2.1 Introduction	8
2.2 Defining academic literacy	8
2.3 The concept of academic reading	9
2.4 Components of academic reading	10
2.5 Vocabulary	10
2.6 Vocabulary Levels Test (VLT)	12
2.7 Vocabulary thresholds	13
2.8 Factors that relate to poor academic reading	14
2.8.1 Age	15
2.8.2 Gender	15
2.8.3 Language/s	16
2.8.4 Poor foundation skills	17
2.8.5 The role of high schools	18
2.8.6 Reading habits	19
2.8.7 Preferred reading format	19
2.9 Approaches to address the reading problem	20
2.10 Conceptual framework	21
2.10.1 BICS/CALP theory	21
2.10.2 Affordance theory	22
2.11 Chapter summary	23
Chapter 3: Methodology	24
3.1 Introduction	24

3.2	Research design	24
3.3	Data collection and variables	24
3.4	Sampling	25
3.5	Data analysis tools and initial processing	26
3.6	Data analysis methods.....	30
3.6.1	Descriptive analysis	30
3.6.2	Bivariate analysis method	31
3.6.3	Multiple linear regression	33
3.7	Ethical considerations	34
3.8	Chapter summary	35
Chapter 4:	Results and Analysis	36
4.1	Univariate analysis results	36
4.1.1	Age and Gender	36
4.1.2	Language/s	37
4.1.3	Reading strategies and reading habits.....	40
4.1.4	Reading formats	42
4.1.5	Word level scores	46
4.2	Bivariate analysis results.....	51
4.3	Multiple linear regression results.....	54
4.4	Chapter summary	56
Chapter 5:	Conclusions and Recommendations	57
5.1	Introduction.....	57
5.2	Summary of main findings.....	57
5.3	Limitations of the study	59
5.4	Recommendations.....	60
5.5	Conclusion	62
Reference List.....		63
Appendix A:	Ethical clearance.....	76
Appendix B:	Diagnostics results - UWL and 5000-word level simple regression	77
Appendix C:	Diagnostics results - Multiple regression	78

List of Figures

Figure 1: Participants' responses to the statements: 'I was taught reading strategies in high school' and 'I enjoy reading'	41
Figure 2: Frequency of time spent reading on cell phones	43
Figure 3: Frequency of time spent reading on laptops.....	44
Figure 4: Frequency of time spent reading on hardcopies	45
Figure 5: Frequency of time spent reading on tablets.....	45
Figure 6: Histogram of 2000-word level results	47
Figure 7: Histogram of 3000-word level results	48
Figure 8: <i>Histogram of 5000-word level results</i>	49
Figure 9: Histogram of UWL results	50

List of Tables

Table 1: Minimums and maximums for the Academic Literacy category, if applying for a degree	9
Table 2: Strategies employed when experiencing difficulties reading	20
Table 3: Selected variables from the RHQ with renamed options	27
Table 4: Selected variables from the VLT with renamed options	28
Table 5: Distribution of students' ages	36
Table 6: Distribution of students' gender	36
Table 7: Frequency of home languages, main languages of instruction in Grades 1 to 3 and in high school among participants	38
Table 8: Frequency of how well students read and write in English	42
Table 9: Vocabulary word level results	46
Table 10: Simple regression model: UWL ~ Reading strategies	51
Table 11: Simple regression model: UWL ~ 2000-word level	52
Table 12: Simple regression model: UWL ~ 3000-word level	53
Table 13: Simple regression model: UWL ~ 5000-word level	54
Table 14: Multiple Regression Model: UWL ~ Reading strategies, 2000-, 3000- and 5000-word level	55

List of Abbreviations

AWL	Academic Word List
BICS	Basic Interpersonal Communication Skills
CALP	Cognitive Academic Language Proficiency
CAPS	Curriculum and Assessment Policy Statements
DBE	Department of Basic Education
DHET	Department of Higher Education and Training
ESL	English as a Second Language
FAL	First Additional Language
NBT	National Benchmark Test
PIRLS	Progress in International Reading and Literary Study
RHQ	Reading Habits Questionnaire
UWL	University Word List
VL	Vocabulary Levels Test

Chapter 1: Introduction - Background and Context

1.1 Introduction

Since the introduction of democracy three decades ago, South African universities continue to see a flood of enrolments from first- and second-generation students, symbolising significant progress in the broadened access to higher education, specifically among non-traditional students who do not come from generational educational backgrounds. While these students enter higher education with much excitement, they usually face challenges, such as difficulties adjusting to university life, academic unpreparedness, lack of commitment, funding and institutional support, which results in them dropping out. The Department of Higher Education and Training [DHET] (2016) reported that the dropout rate for first-year students was 25% in 2005 and by 2014 this cohort's drop-out rate had increased to 38%. One likely reason for the high attrition rate is the well-documented reading crisis in the country (DHET, 2016; Fourie, 2020). Many students are entering higher education with poor literacy levels which leave them unprepared to tackle the demands of tertiary learning (Scheepers, 2016).

In South Africa's dynamic context, it is difficult to underpin one main factor that relates to reading proficiency. To better understand the complexities of reading at tertiary level, this study focuses on the component of academic vocabulary. Vocabulary knowledge influences achievement in reading, writing, and ultimately academic success (Laufer et al., 2004). Therefore, this study investigates the associations between academic vocabulary proficiency and different predictors among South African first-year university students.

1.2 Background and context

Cliff (2024) highlights that the key assessment areas necessary for academic successes include academic literacy, quantitative literacy, mathematics, mathematics comprehension and reasoning. This study focuses on the component of academic literacy. Academic literacy is an integrated set of abilities that is important for critical conversations in academics, teaching and learning curriculum designs and the assessment of students' abilities to apply their language knowledge, in terms of particularly reading and writing (Cliff, 2014, 2024). In an adaptation of Bachman and Palmer (1996), academic literacy encompasses language skills, namely functional, sociolinguistic, grammatical and textual aspects of language knowledge (Cliff, 2014). The concepts of academic literacy and academic reading are interwoven and often

overlapping in their definitions and applications, with academic reading typically recognised as a fundamental component of the multifaceted concept of academic literacy (Kiszcza, 2019).

Andrianatos (2019) argues that reading is a fundamental academic literacy ability required by students in higher education. Several studies have documented the issues related to reading in South Africa and they have found that first-year students in universities lack critical and analytical reading skills necessary for higher education, contributing to poor academic literacy (Andrianatos, 2019; Bharuthram, 2017; Nel et al., 2004; Pretorius, 2000, 2005). By recognising the ongoing concern of poor reading proficiencies in high education, stakeholders can work towards ensuring that students are offered necessary support to help remedy the frustration that may lead to high dropout rates.

This study focuses on vocabulary, particularly academic vocabulary, as a fundamental element of academic reading that falls under grammatical language knowledge (Bachman & Palmer, 1996). Vocabulary knowledge is linked to success in academic reading and writing, which contribute to overall academic success (Laufer et al., 2004). Vocabulary is defined as the reader's ability recognise and work out or derive word meaning from their contexts (Cliff, 2014; Department of Higher Education and Training, 2021). Schmitt (2000) further notes that vocabulary knowledge comprises several components essential for deriving meaning from words, including the breadth of one's vocabulary, knowledge of synonyms, and the ability to appropriately use words within context. The DHET (2021) further argues that unless one understands 95% of the words in a text one is less likely to comprehend what one is reading. Stæhr's (2008) study, which investigated associations between vocabulary size and listening, reading and writing, found statistically significant correlations, with the reading component showing the highest correlation to vocabulary size. Nizonkiza and Van Dyk (2015) also show correlations between academic learning and vocabulary size. Dong et al. (2020) also found relationships between vocabulary knowledge and reading comprehension from foundational phase up to master's level. Related reading skills such as inferencing and comprehension also rely on a rich vocabulary (Weder, 2022). This highlights the valuable role of vocabulary knowledge in supporting proficiency across language skills.

Research within South Africa confirms poor reading levels amongst undergraduate students at various universities. Several studies confirm inadequate reading abilities among undergraduate students, with students struggling to engage with text in meaningful and effective ways (Andrianatos, 2019; Nel et al., 2004; Ngwenya, 2010; Pretorius, 2000; Weder, 2022). The

difference between general and academic vocabulary is that general vocabulary is normal everyday language use whereas academic vocabulary is mainly used in classroom settings, textbooks, texts and assignments (DHET, 2021). The difference between general and academic vocabulary is further highlighted by the concepts of Basic Interpersonal Communicative Skills (BICS) and Cognitive Academic Language Proficiency (CALP), which are further discussed in Section 2.10.1. Research further confirms poor academic vocabulary proficiency among first-year students (Nel & Nel, 2014) which likely prevents students from improving their reading comprehension and general academic literacy and can consequently hinder academic success in higher education.

The problem of poor reading proficiency does not start when a learner sets foot on a university campus; one of the most cited reasons for poor reading skills in South Africa is its multilingual landscape. The South African constitution recognises 12 official languages (Parliament of the Republic Of South Africa, 2023). To promote equality and diversity in education, Section 29(2) of the South African Constitution stipulates that every learner has the right to receive a basic education in the language of his or her choice, where this is reasonably practicable (Stein, 2017). Most students use their home languages in the foundational phase of their education, followed by a sudden shift to English from Grade 4, which proves problematic (Owen-Smith, 2010). UNESCO Institute for Statistics (2025) defines home language (sometimes referred to as first language, native language or mother tongue), as a language learned in childhood, particularly in the home environment.

The constitutional provision for language in education is evidently limited in terms of practicability. Most South Africans are bi- or multilingual, speaking at least two or more languages (Government Communication and Information System [GCIS], 2022). However, multilingual classrooms often make it difficult for learners to use their home languages, the languages with which they are most familiar and comfortable. As a result, they are less likely to perform to the best of their ability (Owen-Smith, 2010). Furthermore, Stein (2017) notes that in bi- or multilingual classrooms, the provincial education department only offers education in a particular language if a minimum threshold is met: at least 40 learners for Grades 1 to 6, or 35 learners for Grades 7 to 12. However, even when this threshold is reached, the implementation still depends on the availability of a qualified teacher who can teach in that language, as well as textbooks for each learning area. With many under-resourced schools in the country, meeting these conditions is a challenge; as a result, learners may be denied

education in their home language, which can hinder comprehension, learner engagement and academic achievement.

Studies highlight that the reasons for poor reading (and vocabulary) levels in South Africa cannot be accredited solely to the switch to English in Grade 4, as the Progress in International Reading and Literary Study (PIRLS) 2021 report on South Africa's reading performance shows that approximately 81% of Grade 4 pupils cannot read for meaning (Department of Basic Education [DBE], 2023), and this includes home languages in which many had instruction until Grade 3 (Mahlangu & Mtshali, 2024). Dagada (2022) notes that, despite the implementation of the National Reading Strategy and the Teaching Reading in the Early Grades interventions aimed at improving learners' reading levels, the teaching of reading in South Africa continues to face significant challenges. These are particularly evident in rural areas and include teachers' limited knowledge of reading methods and approaches, a lack of reading materials, overcrowded classrooms that hinder individualised attention, and insufficient family support, among other issues. Most learners who fail to grasp the reading skills in the foundational primary school level perpetually fall into catch-up mode and longitudinal data shows that they never catch up (Wills et al., 2022). Ngwenya (2010) argues that the issues of the poor reading skills usually start at primary level, persists into high school and ultimately into high education.

It is important to recognise that factors that influence vocabulary proficiency and academic reading are multifaceted, encompassing aspects such as high school performance, issues of race, language, goal-orientation, gender, reading habits, parents' level of education, choice of study, and financial pressure (Dos Reis et al., 2019). In a multifaceted country like South Africa, there are a lot of factors that influence vocabulary proficiency. The factors or predictors to be considered for this study will be selected from the two research instruments considered for this study: the Vocabulary Levels Test (VLT) and the Reading Habits Questionnaire (RHQ); these will be clearly highlighted in Section 3.5. Additionally, the widespread poverty in South Africa, poorly resourced schools and inadequately trained teachers are associated with the poor reading proficiencies in the country (Pretorius & Klapwijk, 2016). Approximately 63% of children in the country are living in poverty. This inhibits the learners' ability to attend school due to the cost of schooling; while many may be motivated to learn they are unable to advance their education (Nortje, 2017). Furthermore, with around 80% of South African schools said to be functioning under poor conditions such as poor funding, lack of resources and unqualified

staff, they are failing to provide students with the necessary skills they require to succeed academically, which further perpetuates the poverty cycle rather than breaking it (Moloi, 2019).

While a lot of research has investigated the reading levels of children in South African primary schools, far less has been done at higher education level. At university level, many lecturers are under the assumption that students can understand most of what they read, and therefore, do not emphasise the teaching of reading (Weder, 2022). This study addresses this gap by focusing specifically on students' vocabulary levels, and the factors specific to this study's identified cohort that relate to academic vocabulary proficiency among first-year students. A better understanding of these predictors will add to the existing body of literature and eventually inform targeted reading-related interventions aimed at enhancing students' reading and academic literacy.

1.3 Statement of problem

In higher education, reading is critical for student success (Andrianatos, 2019). However, studies in South Africa reveal a persistent issue: many students enrol in university with poor reading levels, placing them at a disadvantage and widening academic gaps (Scheepers, 2016). Pretorius (2002) argues that the problems of reading in the country are often masked as a language proficiency issue; while language and reading are linked, this relationship is often oversimplified because, if the problem was this simple, then native language speakers should be good readers in their native languages. The 2021 PIRLS results show poor reading levels in South Africa amongst Grade 4 learners and this includes the home languages in which many students were taught up to Grade 3 (Mahlangu & Mtshali, 2024).

Despite government and research acknowledgment of the reading crisis, there remains a lack of understanding about the multifaceted factors affecting reading proficiency. Huang et al. (2022) argue that vocabulary knowledge plays an important part in reading because unfamiliar words in a text make it difficult for students to comprehend what they are reading. A study by Laufer and Nation (2001) tested students' oral reading fluency across the different Vocabulary Levels Test (VLT) thresholds, and they discovered that fluency and speed correlate to vocabulary size; results for native speakers were similar across all vocabulary thresholds but the speed for non-native speakers lagged as the vocabulary size of target texts increased. The researchers note that they do not know whether the lag of non-native speakers was because of vocabulary or the environment in which the vocabulary was learnt, therefore research on the different learning settings could help with an understanding of vocabulary development. As

further discussed in Section 2.5, there is a lot of research on the influence of vocabulary on reading and overall academic success (Wong & Lee, 2020), but little research investigates what influences vocabulary development and proficiency itself. This study aims to address this gap by examining the relationships between academic vocabulary proficiency and several factors in the VLT and RHQ among first-year university students.

1.4 Research question

To add to the literature on the reading levels of South African students, this study asks the following research question:

Which factors in the VLT and RHQ predict academic vocabulary proficiency among first-year university students?

1.5 Research objectives

To answer the research question above, the study aims to achieve the following objectives:

- * To identify the demographic and behavioural predictors contained in the RHQ and VLT that relate to academic vocabulary proficiency among first-year university students.
- * To explore the associations between academic vocabulary proficiency and each selected predictor among first-year university students.
- * To examine the extent of the relations between academic vocabulary proficiency and the selected predictors among first-year university students.

1.6 Significance of study

Very few studies have attempted to investigate the associations between students' multifactorial contexts in predicting vocabulary proficiencies (Hoang et al., 2021). Rabiner et al. (2016) encourage a need for research that predicts academic achievement from students' competencies rather than the overstated attention on academic outcomes. By identifying some of the different predictors that relate to academic vocabulary proficiency, the study can inform further research on targeted factors or contexts that can empower first-year students to succeed in their academic journey. Furthermore, this study is multidisciplinary, drawing on insights from the fields of languages and literacy, data science, and education, which has an impact on the broader social sciences. This approach creates opportunities for knowledge and insights to be shared across disciplines (Nowell et al., 2022).

1.7 Chapter outline

Chapter 2: Literature Review and Conceptual Framework

Chapter 2 reviews literature on the academic literacy, academic reading, academic vocabulary and factors that relate to academic vocabulary proficiency. This chapter also looks at similar studies that examine academic reading and academic vocabulary and factors that relate to vocabulary proficiency and highlights the conceptual framework of the study.

Chapter 3: Methodology

Chapter 3 describes the methodology in depth and explains the reason for the choices made by the researcher. This chapter comprises of the research design, data collection, data variables, data analysis tool, data analysis method, limitations of study and ethical considerations.

Chapter 4: Results and Analysis

Chapter 4 presents, analyses, and interprets findings on the predictors of academic vocabulary proficiency among first-year university students and the associations between academic vocabulary proficiency and the predictors. The findings are objectively displayed, analysed, and interpreted in alignment with the chosen research methodology.

Chapter 5: Conclusions and Recommendations

Chapter 5 summarises the study's findings and concludes the research by addressing the research question. The chapter discusses the significance of the results and discusses their implications for academic reading. Lastly, based on the findings and limitations, the chapter provides recommendations future research.

1.8 Chapter summary

In conclusion, exploring the predictors of vocabulary proficiency among first-year students is an important investigation with far-reaching implications for student success at tertiary institutions. It is important to understand that the reading challenges are deeply rooted in the country's multifaceted landscape. By researching the predictors of vocabulary proficiency, the study's objective is to yield valuable knowledge into these predictors with the hope to influence further research that can inform evidence-based interventions to support students in higher education.

Chapter 2: Literature Review and Conceptual Framework

2.1 Introduction

This chapter investigates the existing body of literature on academic reading and the predictors of academic vocabulary proficiency among first-year university students. By reviewing the existing literature, this section explores the gaps in research that motivate the need for this study.

2.2 Defining academic literacy

Academic literacy is an important component of student readiness and success at tertiary level. Van Dyk and Van De Poel (2013, p. 56) define academic literacy as the ability to “use, manipulate, and control language and cognitive abilities for specific purposes and in specific contexts.” Academic literacy is considered as the grasp of an array of language skills, such as listening, speaking, reading, writing and reasoning (Andrianatos, 2019). It is difficult to provide a singular definition of academic literacy, given the evolution of the concept due to changes in pedagogy and technology; however, Millin (2015) argues that academic literacy is central to a student’s ability to read and write in English, which is crucial for academic success. These skills are both crucial and closely interconnected, making it difficult to determine which one is more predictive of academic achievement. Millin (2015) argues that reading serves as the primary mode of learning, enabling students to comprehend academic texts, while the ability to write coherent and cohesive texts is essential for university assessment. Although this study focuses on the reading component, with particular emphasis on vocabulary, academic literacy depends on the student’s ability to integrate both reading and writing abilities to succeed academically.

While there are several reasons for student drop-outs at university, unpreparedness to tackle the demands of unguided learning at tertiary level, often linked to poor academic literacy, is one of the reasons often cited (Van Schalkwyk, 2008). In South Africa, the National Benchmark Test (NBT) is a set of assessments commissioned by the Higher Education South Africa (HESA) in 2005, to assess a candidate’s academic readiness for university in three domains of Academic Literacy, Quantitative Literacy and Mathematics (National Benchmark Tests, 2023). The NBT places a candidate’s scores into of three benchmark categories, namely, ‘Proficient’,

‘Intermediate’ and ‘Basic’. Table 1 shows minimums and maximums for the academic literacy category, if applying for a degree. Candidates who fall in the ‘Basic’ category are seen as unprepared for university study and are therefore considered at risk of academic underperformance (National Benchmark Tests, 2024). The 2020 and 2022 NBT results show a slight improvement in academic literacy performance in South Africa, with the proportion of candidates in the Basic band decreasing from 45.48% to 43.01%, respectively (NBT, 2023, p. 67). Despite the modest improvement a substantial gap exists in the preparedness of students entering tertiary education.

Table 1

Minimums and maximums for the Academic Literacy category, if applying for a degree

DEGREE Benchmark Performance Levels	Minimum	Maximum
Proficient	71	100
Intermediate	35	70
Basic	0	34

- From NBT (2024)

2.3 The concept of academic reading

Reading is a fundamental academic task faced by students because students primarily access information this way (Bharuthram, 2012). Reading is defined as “the process of extracting and constructing meaning through interaction and involvement with written language” (Frankel et al., 2016, p. 7). Academic reading can be defined as critical reading that is complex and includes texts that are specific to major subject areas (Sengupta, 2002). Hermida (2009) differentiates between basic reading and academic reading by defining basic reading as a surface approach to reading where the literal meaning of the information is accepted in isolation, whereas academic reading involves an in-depth approach where the student uses higher-order cognitive skills to analyse, problem-solve, and think critically in order to convey the author’s meanings and come up with meaningful interpretations of the text.

In the South African context, a commonly cited reason for high undergraduate dropout rates is inadequate reading abilities (Andrianatos, 2019). Nel et al. (2004) found that on average, first-year students at the University of Potchefstroom were reading at the level of Grade 9 students.

In 2010, Ngwenya conducted a study at North-West University to associate first-year law students' profile with the language in line with their discipline; the study found that on average students scored 48% in a reading comprehension component. For one to comprehend what they are reading, they need to understand 95 percent (95%) of the words in a text (DHET, 2021); therefore, the results are significantly below the mastery threshold. While a lot of recent research exists on reading at primary school level, there is comparatively less literature that provides in-depth recent analysis of university students' academic reading proficiency levels. The NBT, Test of Academic Literacy Levels (TALL) and VLT provide some indication of students' reading abilities; however, these tests provide broad diagnostic overviews rather than detailed insights into the specific factors that challenge students' reading proficiency. Current studies acknowledge the reading crisis and focus on reading habits, strategies and interventions to assist students to succeed academically (Molotja, 2020; Sebola, 2022). This highlights a gap in the research on the factors that impact academic reading proficiency, for example vocabulary.

2.4 Components of academic reading

Reading in general requires an array of skills. The language skills required for academic reading and success can be broken down into four categories: grammatical, textual, functional and sociolinguistic knowledge components (Bachman & Palmer, 1996, pp. 67–68). These language components involve a reader's ability to make essential meaning from text and communicate with intent (functional), recognise cultural nuances and the use of figurative language in different contexts (sociolinguistic), understand word meaning and sentence structures (grammatical), and the ability to understand textual cohesion and organisation beyond the immediate text (textual) (Cliff, 2014). Cliff (2014) points out that there is no research that has attempted a detailed analysis of a single component or subgroup of components of language knowledge needed for first-year students to cope with the pressures of higher education. To address the gap in the literature highlighted by Cliff (2014) and Molotja (2020), this study focuses on vocabulary and factors that relate to students' mastery of various vocabulary levels.

2.5 Vocabulary

Vocabulary is defined as the set of words one must know for effective communication (Neuman & Dwyer, 2009, p. 385). Vocabulary also involves understanding the word's form, deriving its meaning and use from their context (Cliff, 2024; DHET, 2021). Vocabulary, the key focus of

this study (with particular focus on academic vocabulary), falls under the grammatical language knowledge component of academic reading (Bachman & Palmer, 1996). According to the Direct and Indirect Effects model of Reading (DIER), there are several skills and knowledge involved in reading comprehension: word reading, listening comprehension, text reading fluency, background knowledge (content knowledge and discourse knowledge), reading affect or socioemotions, higher order cognitions and regulation (e.g., inference, perspective taking, reasoning, and comprehension monitoring), vocabulary, grammatical (morphosyntactic and syntactic) knowledge, phonology, morphology, orthography, and domain-general cognitions (e.g., working memory and attentional control) (Kim, 2020, p. 469). The model suggests that working memory supports the development of vocabulary and grammatical knowledge, which in turn enables higher-order cognitive skills and self-regulation, ultimately enhancing listening comprehension, text reading fluency, and reading comprehension (Kim, 2020).

Academic vocabulary can be defined as the content-specific words used in subject-areas like biology, geometry, civics, and geography (Baumann & Graves, 2010). Academic vocabulary is typically divided into two categories: general and discipline specific. The general category includes words that occur frequently across academic texts and often have multiple meanings. For example, words like force and function have specific meanings in physics and mathematics but are also used broadly in various contexts (Nagy & Townsend, 2012). In contrast, discipline-specific vocabulary tends to be unique to a particular academic field. Words such as polynomial, cytoplasm, and federalism are typically used in mathematics, science, and history, respectively, and usually carry a single, specialised meaning (Nagy & Townsend, 2012). Vocabulary contributes to reading comprehension both directly and indirectly, as students' awareness of word parts (such as prefixes, suffixes and roots) enhances their understanding of complex word meanings and structures, which in turn improves their ability to construct meaning from text, supports higher-order thinking, and ultimately strengthens reading comprehension (Kim, 2020). Furthermore, as students progress with their academic journey, they are expected to know more words (Laufer & Nation, 1999) and vocabulary size and depth are important for academic reading. The number of words one knows and their ability to derive multiple meaning from a word helps them to comprehend the text they are reading (DHET, 2021). Section 2.6 provides a detailed discussion on the vocabulary word levels that students are expected to master as they progress through their academic journey.

Vocabulary proficiency is linked to language skills proficiency, such as reading, writing, speaking and listening and ultimately academic success (Bayazidi & Saeb, 2017; Laufer et al.,

2004; Molotja, 2020; Neuman & Dwyer, 2009; Wong & Lee, 2020). Kim's (2020) study to investigate the component skills that are directly related to reading comprehension among 201 Korean-speaking first graders found that vocabulary and grammatical knowledge predicted higher order cognitions and regulation, which, in turn, predicted listening comprehension, which lays a foundation for fluent reading and ultimately leads to stronger reading comprehension. A lack of academic vocabulary knowledge in particular has constantly been recognised as a barrier to student success (Nagy & Townsend, 2012). To effectively comprehend written texts, students need to know 95% to 98% words in any given text (Laufer, 2020). Andrianatos' (2019) study shows that students are frustrated because they lack the vocabulary needed to comprehend their course materials textbooks. Molotja's (2020) study on the academic reading strategies of first-year English Education university students, based on their NBT academic literacy test results found that the students' lack academic reading skills, which creates a barrier to accessing information in academic texts and contributes to a higher failure rate.

2.6 Vocabulary Levels Test (VLT)

Laufer et al. (2004) argue that knowing a word develops in levels, beginning with a basic recognition of the word and culminating in the capacity to accurately use it in natural language production. Assessing vocabulary proficiency at university entry is important to help predict students' reading capabilities and determine where educators can intervene (Scheepers, 2016). Paul Nation's (1983) Vocabulary Levels Test (VLT) is one common language test used to assess vocabulary knowledge (Webb et al., 2017). Initially developed in New Zealand at the Victoria University of Wellington, the VLT aimed to help teachers prepare vocabulary teaching and learning plans and has been adapted worldwide to test language skills for English as a Second Language (ESL) learners (Bayazidi & Saeb, 2017). The VLT can measure either receptive (passive) or productive (active) vocabulary knowledge. Receptive knowledge refers to the ability to understand the meaning of a word when encountered in reading or listening, whereas productive knowledge involves the ability to produce an appropriate form or derivation of a known word (Webb et al., 2017). In this study, the productive version of the VLT was used, as we wanted to determine whether students could go beyond recognising a word, by also using it appropriately.

The test categorises vocabulary word levels into 2000-, 3000-, 5000-, University Word List (UWL) and 10000-word levels, with each section comprising eighteen questions (Laufer &

Nation, 1999). A student's vocabulary score is an indication of how many words they know for each level and is important for determining if a student has satisfactory knowledge of a level. Laufer and Nation (1999) suggest that a score of 15 or 16 out of 18, which is around 85% or 90%, is a demonstration of such mastery. Although the VLT is given as a single five-part test, the overall score is less important than the individual scores for each level, which provide more valuable insights (Webb & Sasao, 2013), and this will be important for the analysis in this study.

2.7 Vocabulary thresholds

The 2000- and 3000-word levels represent high-frequency words that appear most often and across the widest variety of language contexts, whether in formal or informal, written or spoken use (Gholaminejad, 2020). These levels are considered foundational vocabulary levels that enable basic communication in English (Scheepers, 2016). Items at the 2000-word level are usually the easiest and tend to have low reliability for both active and passive recognition (Laufer et al., 2004), suggesting that this level does not effectively distinguish learners' vocabulary abilities; whether in recognising words (passive) or using them (active). According to Webb et al. (2017), the 2000-word level is important for basic interpersonal skills. The 3000-word level is the necessary foundational stage for students to begin reading intensive text (Nation, 1983). This level is important when students begin to develop cognitive academic language proficiency (CALP) and transfer language skills learnt in their first language to their second language (Scheepers, 2016). The BICS and CALP concepts will be elaborated further in the conceptual framework subsection (see Section 2.10.1). Knowledge of the 3000-word level is important for progressing to the 5000-word and subsequent levels (Laufer et al., 2004).

The 5000-word level contains mid-frequency words (Gholaminejad, 2020) and it bridges a gap between the final stages of graded reading and the requirement for unassisted comprehension (Nation & Anthony, 2013). This level is important for extensive reading and individual study, which helps individuals to read texts that are difficult (Nation & Anthony, 2013). Knowledge of the 5000-word level allows students to comprehend approximately 98% of most ordinary written texts (Nation & Anthony, 2013).

The 10000-word level is considered the hardest level, and it is categorised as low-frequency words (Laufer et al., 2004). This level comprises rarely used terms; thus, words that are infrequently encountered in language use (Gholaminejad, 2020). To demonstrate the difficulty of the 10000-word level, studies by Laufer and Nation (1999) involving high school learners

from 10 to 12th grades and Kilic (2019) involving Turkish university students between the ages of 18 and 23 excluded the 10000-word level because these researchers did not have learners with sufficient knowledge of the words at that level. Students may recognise words from the 10 000-word level, they are less likely to have sufficient knowledge to demonstrate competency at this level. For the same reasons, the 10000-word level will also be excluded from this study.

For this study, the university-word list (Academic Word List/AWL) level will be considered as the benchmark for academic vocabulary proficiency. The AWL encompasses academic words frequently used in academic discourse across different subject areas (Scheepers, 2016). The AWL covers 570-word families that account for approximately 10% of the total words (tokens) in academic texts (Coxhead, 2000). Items on the AWL are a special case that represent specific domains encompassing academic discourse (Laufer et al., 2004). Knowledge of the AWL can double the number of academic texts that are comprehensible for a student (Gholaminejad, 2020). The AWL serves as an essential measure of students' academic abilities at tertiary level, differentiating academically prepared from underprepared students (Gholaminejad, 2020). Although the AWL is particularly important at the Further Education and Training (FET) and higher education levels, having access to the words on this list during the Senior Phase is beneficial (DHET, 2021).

A study by Laufer et al. (2004) found that items of the AWL could be present at any stage of assessment; thus, at the 2000-, 3000- and 5000-word levels. Knowledge of these levels would only be useful in the AWL level if knowledge of the academic words is relevant to the context in which the words are used; therefore, subject specific context is important. While prior language proficiency is often assumed to impact a student's academic literacy, it is their ability to comprehend their academic texts that holds greater significance in higher education contexts (Cliff, 2014). These arguments stress the nuanced association between vocabulary and academic success at tertiary level.

2.8 Factors that relate to poor academic reading

Dos Reis et al. (2019) argue that there are various factors influencing academic reading, these include gender, reading habits, high school performance, race, language, goal orientation, proximity to school, parents' education, choice of study and financial pressures. While there are several factors that impact academic vocabulary and reading comprehension at large, the next subsections discuss factors that align with the factors assessed by the VLT and RHQ.

2.8.1 Age

While there is no age limit at university, students are classed as ‘young’ if they begin their undergraduate studies at under 21 years and ‘mature’ or ‘older’ if they are aged 21 years and over (Sutton, 2019). Pellizzari and Billari (2012) found that at the undergraduate level there is connection between age and general academic performance; younger students perform better academically than their older peers. Mature students tend to drop out at nearly twice the rate of younger learners, and this disparity may be influenced by life events such as employment, marriage, and childbirth that mature students face but are less prevalent in early adulthood (Sutton, 2019).

Age is often associated with a decline in several cognitive domains, but studies show that there is an exception to certain aspects of language abilities, most notably vocabulary knowledge (Kavé, 2024). Studies that tested the associations between age and vocabulary knowledge found that as one gets older, one’s vocabulary size increases (Kavé, 2024; Keuleers et al., 2015). Kavé’s (2024) study on vocabulary changes in adulthood found that there is ample evidence that older adults perform better on vocabulary tests than young adults do. It is assumed that everyday life represents an opportunity to acquire vocabulary, through education, increased reading and life experiences, and this acquisition is not easily forgotten; therefore, the older one gets, the richer one’s vocabulary (Kavé, 2024; Keuleers et al., 2015). While these studies confirm associations between age and vocabulary, there is little evidence on how age influences academic vocabulary proficiency in particular.

2.8.2 Gender

Rubin et al. (2018) argue that if age differences influence learning, then there is a need to also investigate the influence of gender on learning outcomes. Several studies have investigated the issues of gender disparities in education over the years. Encouragingly, recent statistics in South Africa highlight notable achievements in gender parity at tertiary education, with 13.1% of females and 12.3% males attaining their tertiary qualifications, which is a significant achievement when compared to the 1996 statistics 6.7% females and 8.6% males (Stats SA, 2024). This highlights an upward trend not only in the overall increase of tertiary education, but also a shift in gender dynamics, where the number of female graduates surpasses their male counterparts. Studies shows that female students are more motivated to learn languages compared to their male peers (Montero-SaizAja, 2021; Rubin et al., 2018). Scheepers’ (2016) study on the importance of vocabulary at tertiary level finds that gender is an indicator of

vocabulary knowledge, with women outperforming men. Rubin et al. (2018) also finds that mature female students are more motivated to learn than younger female learners, but generally unable to due to their circumstances as highlighted in Section 2.8.1.

2.8.3 Language/s

Language significantly influences learning. The impact of languages on learning in South Africa has been well-documented. South Africa is a multilingual country, with 12 official languages, namely Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, Afrikaans, English, isiNdebele, isiXhosa, isiZulu and South African Sign Language (SASL) (GCIS, 2022, p. 3). Most South Africans are able to speak two or more official languages (GCIS, 2022). The ability to speak two or more languages does not necessarily mean that students are biliterate or multiliterate; however, Cummins (1979) argues that bilingualism has been positively linked with cognitive and linguistic development. Posel et al. (2020) found that most South Africans speak their African languages inside and outside their homes and the prevalence of the English language in South African homes is minimal. Nonetheless, English continues to be the dominant language in formal settings like schools, businesses and public offices and this dominance has encouraged the growth of English as a second language rather than replacing African languages (Posel & Zeller, 2016).

While students' academic performance is broadly related to literacy, in the South African context, one of the key challenges is their proficiency in English reading (Babane & Maluleke, 2020; Stoffelsma & Spooren, 2019). Most learners in the country speak a different language from the language of instruction in schools starting in Grade 4 (or after Grade 4), namely English (Robertson & Graven, 2020). To rectify the injustices of Apartheid, the South African education policy encourages the use of all official languages in schools and allows school bodies to choose their language/s of instruction (Taylor & Von Fintel, 2016). While this encourages a more accessible learning environment, secondary school leaving examinations in South Africa are only well-developed and administered in English and Afrikaans (Taylor & Von Fintel, 2016). The language autonomy given to School Governing Bodies creates a variation in language policies, which can lead to unequal opportunities for learners, where some schools may choose to introduce English sooner while others only introduce it much later. A study by Van Staden et al. (2016) illustrates that learners from African language backgrounds are particularly disadvantaged when the language of instruction and their home language do

not match and reading literacy achievement scores tend to decrease significantly when this misalignment exists.

Babane and Maluleke (2020) argue that multilingualism should be encouraged at South African universities to create a comfortable environment that enables students to actively participate in their studies and deepen their language comprehension, which will improve their academic performance. The University of Western Cape started an initiative where course materials are translated using other official languages and Dyers and Antia (2019) found that students felt that the localised translations offered better access to the course content. A study by Wong and Lee (2020, p. 72) found it is more challenging to produce a text in a second language because ESL learners spend a considerable amount of time and effort composing, thinking, and investigating ideas in the target language. In the South African context, Khanyile and Awung (2023) highlight that there are significant limitations to teaching in indigenous languages and this includes limited subject-specific terminology and a lack of African written languages resources, which raises quality concerns and possible risks of miscommunication. Therefore, languages play an important role in academic success.

2.8.4 Poor foundation skills

The impact of foundational learning on tertiary level education is often understated. Most learners in South Africa complete their three years of foundation phase in indigenous languages, whereafter they switch to English as they advance to Grade 4 (Owen-Smith, 2010); this at least partly results in poor English reading skills from Grade 4 learners. However, Weder (2022) cautions that the challenges facing South African students are not simply because of second language instruction; rather, they reveal a lack of foundational reading skills. PIRLS reports show that a vast majority of South African Grade 4 pupils cannot read for meaning in any language, including the African languages they are taught in at Grade 1 to 3 (Howie et al., 2017; Mahlangu & Mtshali, 2024). The PIRLS report shows that 78% of Grade 4 learners could not read for meaning in 2016 and the problem worsened during the COVID pandemic as highlighted by the 81% Grade 4 learners who could not read for meaning reported in 2021 (DBE, 2023). While the PIRLS mainly assesses reading comprehension rather than foundational reading skills, these results suggest that students may not have acquired the skills necessary for learning to read needed to support reading for meaning.

Most learners who fail to grasp the foundational reading skills at the primary level perpetually fall into catch-up mode, and longitudinal data shows that they never catch up (Wills et al.,

2022). It is generally understood that children with poor reading proficiencies are at high risk of developing behavioural and emotional disorders and it reduces their chances for future success like advancing in education, work, and social well-being (Daniel et al., 2006). The ability to read for meaning at an early age prepares children with the foundational learning skills they need to read and make evidence-based interpretations at later stages in their education and lives.

2.8.5 The role of high schools

Investigations into the reading skills of high school learners is also overlooked (Weder, 2022). High school teachers play an important role in exposing students to effective reading strategies necessary to equip them with skills needed for reading comprehension and to help them succeed in their academic journeys (Matope & Senye-Awudi, 2024). Nel et al. (2004) argue that many first-year university students underwent a high school experience that left them underprepared for the demands of university learning. Students who lack reading skills and vocabulary knowledge from high school often face more difficulties reading texts in university which results in high university drop-out rates (Bharuthram, 2012). In high school, students are expected to transition from teacher-guided reading to unaided reading that enables them to tackle various reading sources (Matope & Senye-Awudi, 2024). This is also the stage when students are expected to master the 5000-word level threshold discussed in Section 2.7. This, together with knowledge of appropriate reading strategies, will equip them with necessary reading skills and assist them when faced with dense university texts.

In a study with Grade 8 high school learners to test their reading abilities in Setswana and English, Matjila and Pretorius (2004) found that, learners were reading at a Grade 3 to 4 level, four years below their expected level. This gap is likely to persist as students continue with their studies. Sebolai (2022) notes that a gap exists between the strategies students leave high school with and the required knowledge for tertiary education for students to succeed academically. Students struggle with transferring specific strategies learned from high school to the academic literacy demands of each course (Nel et al., 2004). A study by Dos Reis et al. (2019) on two high schools in the Western Cape Province found that the lack of communication between high schools and universities on the expected foundational knowledge and skills worsens the reading problem.

2.8.6 Reading habits

While the necessary reading techniques may be taught to the students at high school level, it is equally their responsibility to sharpen these skills by doing more reading on their own (Romadlon, 2017). A reading culture offers holistic benefits where students are not only entertained but they can also expand their vocabulary and improve their language skills. Through reading, learners are exposed to new words and learn how to connect them to other words, thereby deepening their vocabulary knowledge. The more students read, the greater their exposure to vocabulary that is rarely encountered in spoken language; what is more, written texts expose learners to complex linguistic structures, and complex subordinate clauses that further enrich their language development (Mart, 2012). Bharuthram (2017) argues that increased reading extends beyond knowledge acquisition; it aids language skills transfer and has the potential to improve a reader's attitude towards reading. While reading academic text is different from reading for pleasure, studies show that students with poor reading habits often face challenges when reading other types of texts (Anwar & Sailuddin, 2022). Bharuthram (2017) found that some students are concerned that if they read for leisure, they will not have time to focus on their demanding academic work. This concern reflects a perception among students that leisure reading is a distraction, and has a lower priority compared to assigned academic tasks.

2.8.7 Preferred reading format

Technological advancements in the education sphere have influenced an expansion of concepts such as text, reading, and literacy (Mangen & Van Der Weel, 2016), with the expansion from print to digital formats providing students with a wider range of reading materials and format options. The 2022 census found that 92.1% of households in South Africa owned a cell phone with 60.5% of the population accessing internet mainly through their cell phones (South African Government News Agency, 2023). The popularity of cell phones in households can be motivated by the issue of portability and/or affordability and understanding how this influences students' reading habits is fundamental.

Margolin et al. (2013) investigated the effect of using electronic media or paper on reading comprehension and found that there was a lack of significant differences in comprehension accuracy across the different media platforms. Some studies have found that digital devices often lead to distractions, as students may spend time navigating how to use the device or

browsing unrelated content. To ensure effective comprehension during reading electronic media, strategies are needed to minimise these distractions (Spjeldnæs & Karlsen, 2024).

2.9 Approaches to address the reading problem

Addressing poor academic literacy at university level requires a multifaceted approach that targets skills development, student backgrounds and support systems. Nel and Nel (2014) urge that institutions must actively address issues of literacy development; measures need to be implemented to ensure academic reading proficiency, and this includes supporting students' vocabulary acquisition. Reading comprehension is not emphasised as much as writing is during university-level interventions. Nonetheless, universities have an ethical obligation to implement strategies that support students in overcoming barriers to their success, and one such strategy involves thoughtfully designed academic literacy interventions (Fouché, 2024). One of the shortcomings for academic literacy courses is that they are usually generic and do not accommodate each student's special needs (Andrianatos, 2019).

A student's academic journey is equally their responsibility, and it is important for students who are struggling to seek help. Findings by Bharuthram (2017) in Table 2 show that some students do admit that they are experiencing reading difficulties, and they employ different strategies to address the problem.

Table 2

Strategies employed when experiencing difficulties reading

Reading strategies used	Percentage (n=62)
- Seek assistance from friends	75.08%
- Seek assistance from lecturers	09.67%
- Use a dictionary to understand unknow words	51.61%
- Read the text more than once	45.16%

- From Bharuthram (2017, p. 54)

Table 2 shows that most students are more comfortable with seeking assistance from their friends than they are with asking their lecturers. Kruger & Dunning's (1999) theory, the Dunning-Kruger effect, argues that individuals with low knowledge or lack expertise in a particular area do not recognise their own limitations and as a result they overestimate their

competences. Andrianatos (2019) further highlights that despite the inadequate reading abilities, there is a cycle of non-compliance amongst students, where they do not allow enough time to read and engage with a text. Only 45% of students read a text more than once to help mitigate the problem (Bharuthram, 2017). Therefore, one might ask if relying on friends who may not realise their shortcomings and also be non-compliant readers would be the best solution to the reading problem.

2.10 Conceptual framework

This subsection outlines the conceptual framework that has been selected for this study. A conceptual framework describes researcher's understanding of the study's variables, their relationships to one another and links to relevant literature (Luft et al., 2022). In this section, underlying assumptions and guiding principles that shape the research will be discussed, offering insights into how these theories interact to address the study's objectives. The study employs two theories towards its conceptual framework, namely, the BICS/CALP theory, and the Affordance Theory.

2.10.1 BICS/CALP theory

The Basic Interpersonal Communicative Skills (BICS) and Cognitive Academic Language Proficiency (CALP) theory was introduced by Cummins in the early 1980s to draw attention to challenges encountered by second language learners in academic aspects, with particular attention to school language (Cummins, 2008). BICS refers to conversational fluency in a language, typically seen in a child's early language development, and involves an embedded context where gestures, tone, and the conversational setting contribute to meaning, while CALP refers to students' ability to understand and express themselves in more decontextualised contexts where meaning must be explicitly framed, which is essential for academic success (Cummins, 2008). Matjila and Pretorius (2004) note that children acquire BICS in their home language before they begin school; they then start acquiring CALP in school when they learn to read and write. Cummins (2008) indicates that CALP can be distinguished from BICS after the initial stages of school, reflecting largely on the language children need for academic success. In today's world, one can develop CALP without being literate because academic language is inevitably social, and involves everyday language (BICS); however, understanding academic language without formal literacy will be limited as CALP demands the student to understand meanings in more decontextualised contexts, constructing meaning through language within the text itself (Aukerman, 2007).

Most learners in South Africa are at a disadvantage because their English CALP journey is delayed because English is not their home language, yet English is the main academic language at universities (Van Der Merwe, 2018). Sibanda (2017) highlights that the BICS/CALP theory brings to light the unequivocal demands placed on English as an Additional Language (EAL) learners in bilingual contexts that are not expected for home language learners in monolingual settings. This suggests that in bilingual contexts, EAL students must develop both every day and academic language skills in additional language, which places extra demands on them, whereas home language students in monolingual settings do not have this additional challenge, as they learn and use their first language both in social and academic contexts.

2.10.2 Affordance theory

Acknowledging that there may be other factors beyond language skills that relate to academic literacy, the influence of the Affordance Theory on how the environment influences learning is also important. The notion of the Affordance Theory was initially formulated by ecological psychologist James Gibson in 1977, describing that there are properties that exist in the environment that enable action potentials (Moll et al., 2022). In the context of education, the Affordance Theory acknowledges that learning may be influenced by a student's environment. Findings from Hadjerrouit and Nnagbo's study (2021), which explored students' and teachers' perceptions of affordances and their actualisation while interacting with the e-assessment system Numbas, found that many affordances emerged in the learning environment, particularly technological affordances, such as the instant feedback provided by computers, students were motivated to engage with Numbas. While some affordances were not realised, those that were actualised sometimes failed to materialise fully due to other constraints, for example concerns about the incorrect marking of some correct answers, which compromises the reliability of feedback (Hadjerrouit & Nnagbo, 2021). Moll et al. (2022) argue that it is important to seek specific pedagogical affordances that can aid academic success. However, in the dynamic South African landscape, it is difficult to underpin one main factor that relates to academic vocabulary proficiency. It is, therefore, important to study some of these affordances of the learning environment that influence the levels of academic vocabulary proficiency and their relationship. With relation to vocabulary proficiency, affordances such as abundant reading materials, different reading formats, and quality reading pedagogy provide learners with opportunities to develop their vocabulary and ultimately, influencing their proficiency and reading comprehension.

2.11 Chapter summary

In the light of the above literature review, there are few studies that investigate the factors that relate to academic vocabulary proficiency and as more students continue to enter higher education institutions with below-par levels of academic reading, this warrants for a study that explores some factors that influence academic vocabulary proficiency among first-year students.

Chapter 3: Methodology

3.1 Introduction

The methodology chapter highlights the researcher's choices on how the study was conducted. This section includes details on the research design, sample, setting, data collection, variables, analysis techniques, limitations of the study and ethical considerations. These steps and decisions are fundamental to systematically work towards addressing the research question and achieving the research objectives.

3.2 Research design

For this study, an exploratory research design was employed to explore the association between academic vocabulary proficiency and the selected predictors among first-year university students. Exploratory research seeks to gain insights into a particular phenomenon by examining and describing the characteristics of variables whose nature may not yet be well understood (Singh, 2021). The exploratory research design is employed to help the researcher develop initial ideas and insights that influence the direction of any further research (Olawale et al., 2023). The study uses quantitative research methods that emphasise the statistical analysis of data collected using computational techniques (Babbie, 2010). By drawing on exploratory research, the study offers a robust and flexible approach in attempting to investigate the predictors that have associations with academic vocabulary proficiency. Given the limited knowledge on reading and English vocabulary levels of university students in South Africa, an exploratory research design is appropriate to guide and inform future studies in this area.

3.3 Data collection and variables

The study uses existing anonymised data collected from 1) vocabulary test results and 2) responses from a Reading Habits Questionnaire (RHQ) that were administered online in 2023 and 2024 among first-year students at the University of Witwatersrand's (Wits) School of Education. The VLT comprises 2000-, 3000-, 5000-, university word list (or Academic Word List), 10000-word levels and total score, and each section has eighteen questions (Laufer & Nation, 1999). Bayazidi and Saeb (2017) argue that the validity of the VLT is based on the idea that instead of focusing on the test's total score, the VLT measures knowledge of words at each level, offering a comprehensive and nuanced understanding of a person's vocabulary proficiency. The VLT also included information on the students' age, gender and home

language. The UWL from the VLT was selected as the dependent variable (y) and serves as the benchmark for academic vocabulary proficiency.

The RHQ is comprised of semi-structured questions, featuring 23 main questions that expand a total of 122 sub-questions for individual response items. From the RHQ, possible variables chosen for exploration include questions on the perception of how well students believe they read and write in English, language of instruction in Grades 1 to 3 and high school, reading techniques taught in high school, reading habits, and students' preferred reading format. Tables 2 and 3 in Section 3.5 provide the detailed list of variables that have been selected for the study.

The datasets consist of continuous and categorical data that were analysed. Continuous data is quantitative data that can be counted in decimals and does not have gaps between data points, and deals with objective measurements (Ranganathan & Gogtay, 2019). For this study, the word-level test scores from the VLT are the continuous data. Categorical data represents that data that can be grouped into subtypes; this data is classified depending on the number of categories (Ranganathan & Gogtay, 2019). All the data from the RHQ, as well as the information on age, gender, and home language collected through the VLT, is categorical. The data from the VLT and RHQ were combined to form one dataset for analysis. Combining information from different datasets greatly enhances results and accuracy, especially when integrating small datasets that contain diverse types of data (Nguyen et al., 2023). Therefore, the decision to combine the two datasets, the VLT and RHQ, is driven by the aim to compare students' perceptions with their actual competencies.

3.4 Sampling

In 2023, 212 participants completed the RHQ and 285 completed the VLT. While 121 students agreed to participate in the study, only 71 of those participants completed both the RHQ and VLT. In 2024, 148 participants completed the RHQ, and 383 completed the VLT. Of the 100 students who agreed to participate in the study, only 62 participants completed both the RHQ and VLT. The total number of participants for this study was therefore 133 ($n = 133$). The reduced sample size may be due to some students not providing their details on either the VLT or RHQ which made it difficult to match their responses across data collection instruments. Some students completed the RHQ twice, possibly misinterpreting the instructions to complete the RHQ and VLT as a directive to complete the same questionnaire twice. Additionally, the

lower response rate on the RHQ in 2024 could be attributed to a shortage of tutors to facilitate the process.

3.5 Data analysis tools and initial processing

The data analysis tool R was used to analyse the data. R is a powerful, open-source programming language for statistical analysis and visualisation, featuring excellent graphics and built-in functions, probability distributions, and simulations (Nolan, 2012). Rousseaux et al. (2013) agree that R is a suitable analysis tool, particularly for social science researchers to efficiently manage survey data, enable secure storage, detailed documentation, data exploration, recoding, and seamless sharing of datasets. Datasets in the form of Excel spreadsheets titled 'Pre-test Vocab (2023)', 'RHQ (2023)', 'Pre-test Vocab (2024)', and 'RHQ (2024)' were carefully reviewed prior to uploading them to R to confirm data accuracy, completeness, and consistency across files. There were no missing values (NAs); therefore, none of the participants were dropped. After confirming the sample sizes, the datasets were merged and stacked for analysis. The 'head', 'tail' and 'summary' functions were used to help check the first few rows (head), last few rows (tail) and overall summary (summary) of the dataset. The results confirm that everything was in order before proceeding with further analysis.

The original variable names were too lengthy, so the researcher renamed them to make the coding easier. Table 3 shows the renaming for RHQ variables. While the names for the word level variables were precise, the rules of syntax in programming languages state that a variable cannot start with numbers (Severance, 2024), as starting with a number confuses a programming language assuming the number is a syntax. R does not recognise variables that start with numbers. Therefore, to make the variables R-friendly, the predictors were renamed (see Table 3).

For this study, the UWL has been selected as the dependent variable (y) that represents the benchmark for academic vocabulary proficiency. The independent variables or predictors (x) considered for analysis are age, gender, home language, how well students read and write in English, the main language/s of instruction at Grade 1 to 3, the main language/s of instruction at high school, how much students enjoy reading, whether students were taught reading strategies at high school, how much time students spend reading on cell phones, laptops, tablets and hard copies, 2000-, 3000- and 5000-word levels.

Table 3*Selected variables from the RHQ with renamed options*

Original names	Shortened variable names
Student Numbers	<i>student_ID</i>
How well do you read and write in each of these languages? [English]	<i>English_readwrite</i>
What was the main language of instruction at your school in Grades 1 to 3?	<i>instructionlang_gr1to3</i>
What was your main language of instruction in high school?	<i>instructionlang_highschool</i>
Please rate the following statement in terms of how much you agree or disagree with it. [I enjoy reading.]	<i>enjoyreading_rating</i>
Please rate the following statement in terms of how much you agree or disagree with it. [I was taught reading strategies in high school.]	<i>reading_strategies</i>
How much time do you spend reading in each of these formats? [Cellphone]/	<i>cellphone_readingtime</i>
How much time do you spend reading for academic work in each of these formats? [Cellphone]	
How much time do you spend reading in each of these formats? [Laptop]/ How much time do you spend reading for academic work in each of these formats? [Laptop]	<i>laptop_readingtime</i>
How much time do you spend reading in each of these formats? [Tablet]/ How much time do you spend reading for academic work in each of these formats? [Tablet]	<i>tablet_readingtime</i>
How much time do you spend reading in each of these formats? [Hard-copy format]/ How much time do you spend reading for academic work in each of these formats? [Hard-copy format]	<i>hardcopy_readingtime</i>

Table 4*Selected variables from the VLT with renamed options*

Original names	Shortened variable names
Year	<i>year</i>
Age	<i>age</i>
Gender	<i>gender</i>
Home Language	<i>home_language</i>
2000-word level	<i>wordlevel2000</i>
3000-word level	<i>wordlevel3000</i>
5000-word level	<i>wordlevel5000</i>
University-word level	<i>UWL</i>

It is often assumed that respondents can complete the questions with absolute certainty without any errors; however, there are instances where there are respondent uncertainties; therefore there is a need to address these (Lundhede et al., 2009). There are instances in this study where students selected several responses for one question, thus creating new response categories. Lundhede et.al. (2009, p. 121) suggest that there are three different ways of handling uncertain responses, namely that they be: “i) eliminated from the sample, ii) asymmetrically recoded in the sense that a choice reported to be uncertain is recoded as a choice of the status quo alternative, or iii) symmetrically recoded such that an uncertain choice is recoded as a choice of the best alternative different from the one chosen.” Lundhede et.al. (2009) further argue that addressing ambiguous responses largely relies on the researcher’s interpretation and objectives of the study. This flexibility allows researchers to adopt an approach that aligns with their methodological framework and the goals they aim to achieve.

While there is no way to tell which response students intended to select, in this study, the first responses were chosen as the best alternative. Therefore, the researcher decided to recode the categories to the available levels instead of removing the participants to preserve the sample size. For example, the researcher standardised the spellings for the language names to align with official spellings to prevent inflating the number of language categories by treating variations as separate entries, which creates inaccuracies. Variants like “Siswati” were thus recoded to “siSwati” and “Zulu” to “isiZulu”, among others. There are also instances where students chose two or more responses, when they were meant to select only one response, thus creating variations. For example, thirty-six students selected two or more languages as their

main language of instruction in Grades 1 to 3 and in high school. These responses were recoded as '*bi-multi_lingual_w_Eng*' for responses that include English and '*bi-multi-lingual_w/o_Eng*' for those excluding English.

The RHQ asks how much time students spend reading on different reading formats, namely cell phones, laptops, tablets and hardcopies. There are students who chose both 'Three to four hours per week' and 'Five to ten hours per week' options, creating a new response category 'Three to four hours per week, Five to ten hours per week'. The first option in this instance was chosen as the best alternative. Additionally, the researchers who developed the questionnaire realised after the 2023 cohort had completed the questionnaire that they had neglected to add a "None" option, which is a significant shortcoming in the 2023 questionnaire. To rectify this, the 2024 questionnaire had the 'None' option. One could argue that this option gave the 2024 cohort an explicit choice that indicates the absence of a particular preference, while the 2023 cohort was forced to choose an alternative response that might not accurately reflect their preference. Furthermore, the 2023 RQH asked students '*How much time do you spend **reading** in each of these formats?*', while the 2024 questionnaire asked, '*How much time do you spend **reading for academic work** in each of these formats?*'. There is a difference between the questions as one asks them about reading in general and the other explicitly asks about academic reading. Therefore, due to these differences, the reading formats were analysed separately using the different years but unfortunately the variables were not included in the bivariate analysis and multiple regression.

Multiple response categories often have low numbers of responses for some categories and in some cases the researcher is motivated to collapse low frequency categories into the other categories to reduce sparse data. DiStefano et al. (2021) recommend collapsing low cell frequency at a minimum of 5 % - 10%, with a higher frequency recommended when the sample size is small; this will help improve the accuracy of the estimates and reduce bias. For example, the languages and reading and writing rating variables had several low frequency responses and to yield meaningful results the researcher also decided to recode the low frequency responses. Therefore, any cell frequency of less than 10% (< 10%) in this study were collapsed into the 'Other...' categories.

3.6 Data analysis methods

The study aimed to identify predictor(s) with statistically significant associations with the UWL. To achieve this, a multiple regression analysis was conducted, with univariate and bivariate analyses serving as the preliminary steps building towards the main analysis (Bertani et al., 2018).

There are sample size requirements for multiple regression to achieve the desired statistical power and decrease error (Van Voorhis & Morgan, 2007). A larger sample size is often recommended but there are practical realities in conducting research that can limit the sample size as highlighted in Section 3.4. There are different versions of rules around the sample size requirements. The commonly used ‘rule of thumb’ for a multiple regression is a minimum of 10 participants per predictor (10:1 ratio) (Van Voorhis & Morgan, 2007; Wampold & Freund, 1987). In total, there are 133 participants ($n = 133$) and 16 predictors for analysis.

$$\text{Sample size estimate} = 10 \times 16 = 160$$

Using the 10:1 ‘rule of thumb’, this study’s sample size ($n=133$) is smaller than the recommended sample size. There are several ways one may address this shortfall, like reducing the number of predictors or using a more lenient ratio. Hair et al. (2010) suggest a more lenient ratio of 5:1, meaning the ratio should never fall below five observations for one predictor. Alternative sample size approaches were only considered when the analysis progressed to multiple regression.

3.6.1 Descriptive analysis

A univariate analysis was employed to fulfil the first objective, namely, to identify predictors that relate to academic vocabulary proficiency. A univariate analysis involves a descriptive analysis of one variable at a time to describe its distribution (Canova et al., 2017). The univariate analysis of numerical values provides details on the minimum, maximum, median, mean, first and third quantile (Rousseaux et al., 2013). The univariate analysis of the categorical data uses the response categories to report on the frequency or proportion of each level within the dataset.

The univariate analysis for the word levels summarises the results which shows the minimum, first (1st) quartile, median, mean, third (3rd) quartile, maximum values and the number of students who achieved the 85% mastery threshold and above. Histograms are plotted to visually

represent the data and their distributions. A histogram is a graphic that shows the number of objects that appear in a certain class interval and helps the researcher to know the distributions of the variables (Li et al., 2019). The histograms show three types of distributions: normal, left-skewed and right-skewed distributions. The distribution of data is significant in the choice of statistical test for analysis to be employed in a study (Ranganathan & Gogtay, 2019). A normal distribution is a symmetric bell-shaped curve that shows that most of the data observations are centred around the middle and the curve gradually tapers on either side (Ranganathan & Gogtay, 2019). When a variable is left-skewed, its tail is heavier on the left side than it is on the right, whereas when it is right-skewed, the right tail is heavier it is on the left (Campisi et al., 2023).

3.6.2 Bivariate analysis method

Once the predictors were identified, the next step was to conduct a bivariate analysis with the objective of exploring the associations between the UWL and the selected predictors of the academic vocabulary proficiency amongst first-year students. Bivariate methods involve analysing the associations between two variables without considering how other variables influence this association (Bertani et al., 2018). A simple regression is the method employed for the bivariate analysis. Hair et al. (2010) define a simple regression as a regression model that involves the dependent variable and a single independent variable. The equation for a simple regression is as follows:

$$Y = (bX + a) + e$$

Where Y is the dependent variable, X is the independent variable, b is the regression coefficient predicting Y from x , a is the intercept constant and e is the error term that captures the unexplained variation in the model (Wampold & Freund, 1987). In a regression, a p -value of less than the 0.05 ($p < 0.05$) confirms if the relationship is statistically significant (Singh, 2021), and the results are interpreted using the coefficients. A coefficient indicates the direction of the relationship between the two variables, which can be negative or positive (Hair et al., 2010).

The data consists of continuous and categorical data. The word levels scores are the continuous data, and the rest of the variables are categorical. In a simple regression that involves non-numeric values (categories), a category is omitted and this acts as a point of reference for interpreting the other categories (Hair et al., 2010). The computational tool can choose a default

reference group or the researcher can make the choice based on either the category that has the largest sample size or the one that makes sense to compare the others to base on the aim of the study (Temko, 2024). For this study, the largest group was used as the reference category; however, for the ‘main language of instruction at Grade 1 to 3’ variable the ‘other languages’ category was the largest, but this was a mix of languages, therefore, the ‘English’ was selected as the most sensible category.

Diagnostics are an integral part of a regression analysis. These are used to investigate if the model’s assumptions are consistent with the data; this is achieved by identifying extreme points or outliers that may be influencing the regression and possibly in the results (Ferré, 2009). Tests of residuals can be used to identify any problem areas in the regressions. Hair et al. (2010, p. 34) define residuals as unexplained portions of the dependent variable in a regression. Graphical and numerical techniques can be employed to identify outliers and influential data points in regression analysis (Chalton & Troskie, 1992). Skewness is another way to gain an understanding of the direction of the irregularities in the data (Campisi et al., 2023). This study employs four diagnostic methods to test for the existence of outliers or skewedness, namely studentised residuals, Cook’s Distance, quantile-quantile plots (Q-Q plots) and histogram of residuals.

Histograms have been discussed in detail in Section 3.6.1. The studentised residuals locates outliers by omitting or deleting the i th observation in the calculation of the regression estimates and each time refitting the regression model on the remaining $n-1$ observations (Hair et al., 2010). Any studentised residual observations that is larger than three (-3 or 3) is considered an outlier (Pennsylvania State University, 2018). To further confirm the diagnostics, the Cook’s Distance is employed as a way to measure influential points in a regression by looking at the difference between regression parameter estimates and the estimates obtained when the i th data point is deleted (McDonald, 2002). A practical rule of thumb for influential points in the Cook’s Distance is that values greater than 1 (i.e., >1) are considered influential (Chatterjee & Hadi, 2012).

Q-Q plots were further employed to confirm any extreme data points. Q-Q plots are a statistical tool used to visualise a sample’s distribution and understand the extent and type of deviation of the variables (Weine et al., 2023). To interpret the results of a Q-Q plot, one should examine how the points are distributed relative to the line (Loy et al., 2015). The line is used as a reference that helps to detect any departures from the proposed model (Almeida et al., 2019).

Points that lie on or close to the line suggest a normal distribution, while points that deviate from the line may indicate potential skewness. Ferré (2009) urges that once outliers have been identified, it is good practice to investigate the possible reasons for this behaviour and decide how they would address the outlier. One can correct the value if there is a mistake, keep the value as it is, or remove it from the regression to see how the regression performs without it (Ferré, 2009).

3.6.3 Multiple linear regression

The final objective of analysis was to examine the extent of the relations between academic vocabulary proficiency and the bivariate statistically significant predictors among first-year university students, when considering other factors. Bertani et al. (2018) argue that the world is “multi variated”; therefore, accrediting associations to any relationship between two variables without considering the impact of other factors results in weak attributions. Therefore, a multiple linear regression can help to formulate a better understanding of the relationship between the variables. Hair et al. (2010) define a multiple regression as an analysis method that involves one dependent variable presumed to be related to two or more predictors. By including and acknowledging other factors, the study recognises real-life relationships and that there are many factors that contribute to the result; this also minimises the influence of the other factors affecting the correlation (Goundar, 2012).

The equation for a multiple regression is:

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \dots + B_nX_n + e$$

In the equation, Y represents the predicted value of the dependent, B_0 the intercept or constant value, X_s are different predictors, and n is the number of predictors (Plonsky & Ghanbar, 2018).

The sample size calculations in Section 3.6 highlighted that the sample size fell short of the requirements for a multiple linear regression (27 participants short). However, during the descriptive and bivariate analyses some variables had to be left out before proceeding with the multiple regression. In the univariate analysis, it was found that the reading formats questions were framed differently in the 2023 and 2024 questionnaires as highlighted in Section 3.6.1; therefore, the variables were left out from the simple and multiple regression. During the simple regression, several variables were not statistically significant and were therefore not included

in the multiple regression. Only four variables were statistically significant; hence, the sample size calculations were revised as follows:

$$\text{Sample size estimate} = 10 \times 4 = 40$$

With 133 participants and four predictors, the sample size exceeds the recommended sample size requirement for a multiple regression analysis, which provides a strong foundation for reliable estimates and the power to detect meaningful associations.

The same diagnostic tests used in the simple regression, studentised residuals, Cook's Distance, Q-Q plots and histogram of residuals, were used to diagnose the multiple regression. These diagnostic tests are discussed in Section 3.6.2. A multiple regression also has the possible risk of multicollinearity. Multicollinearity is when one independent variable is highly correlated with another independent variable, this makes it difficult to determine the distinctive contribution of each predictor to the outcome variable (Hair et al., 2010). Hair et al. (2010) suggest three tests for multicollinearity, namely Correlation Matrix, Tolerance and Variance Inflation Factor (VIF). However, the Correlation Matrix and Tolerance do not accept tests for non-numerical values, thus the VIF was used for this study. VIF indicates which coefficients represent a problem of multicollinearity in the model (Ferré, 2009). A rule of thumb for the VIF is the result should be greater than one (>1) but less than 10 (<10) (Ferré, 2009; Miles, 2014).

Once the diagnostic tests and VIF confirm that there are no issues of outliers and multicollinearity, there are other statistical insights that are important in interpreting the model. The R-squared of a multiple regression has extreme importance in model interpretation, indicating the proportion of variance in the dependent variable explained by the predictors (Wampold & Freund, 1987). The adjusted R-squared is mostly considered as it takes into account the sample size and model efficiency (Plonsky & Ghanbar, 2018). To ensure the model is not overfitting, the adjusted r-squared was also interpreted.

3.7 Ethical considerations

Ethical clearance for this research was obtained from Wits School of Social Sciences (Ethics clearance/waiver number: WSOSSINTR092024/7). One main ethical consideration for this study is anonymity. Students were invited to consent if they wanted to be a part of the study on condition that their names and student numbers will not be revealed to anyone. Therefore, to

honour this condition, the research re-used existing data that had already been anonymised. Furthermore, the cohorts are reported on collectively, and no individuals are discussed specifically. Therefore, no students are identifiable in any way. By adhering to these ethical considerations, the researcher has conducted the study in a manner that respects the rights and well-being of participants while upholding the principles of research integrity and ethical conduct.

3.8 Chapter summary

This exploratory study utilises data collected from the VLT and RHQ questionnaires to investigate associations between academic vocabulary proficiency and the selected predictors. The analysis includes univariate, bivariate, and multiple regression techniques used to investigate the associations and address the research objectives. The data analysis process highlights the researchers' decisions to eliminate inconsistencies and ensure thorough results. Ethical considerations of informed consent and anonymity were upheld throughout the research process by ensuring that results are were discussed collectively.

Chapter 4: Results and Analysis

4.1 Univariate analysis results

The first objective of the study is to identify the predictors that relate to academic vocabulary proficiency among first-year university students. The descriptive analysis results are provided in the subsections that follow.

4.1.1 Age and Gender

Table 5 shows that most students were between the ages of 18 and 19 years old, followed by those aged 16 and 17, 20 and 21, and a smaller group who were 22 years and older. The higher number of students in the 18 to 19-year age group is likely because students enrol in university immediately after high school.

Table 5

Distribution of students' ages

Age	Frequency
16 – 17	17% (<i>n</i> = 23)
18 – 19	62% (<i>n</i> = 83)
20 – 21	16% (<i>n</i> = 21)
22+	5% (<i>n</i> = 6)

There are more female students compared to males (see Table 6), highlighting a significant gender disparity among the participants. Historically in South Africa, tertiary education has been dominated by males; however, enrolment numbers of female students, particularly black female students, have steadily increased over the years (Msimango & Motala, 2023).

Table 6

Distribution of students' gender

Gender	Female	Male
Frequency	69% (<i>n</i> = 92)	31% (<i>n</i> = 41)

Broekhuizen and Spaull (2017) argue that the changing gender norms and rising expectations on of the labour force could explain the increase of female participation in tertiary institutions. Additionally, the participants are from the School of Education, particularly those pursuing the BEd degree and United Nations Educational, Scientific and Cultural Organisation (2023) highlights that teaching is increasingly a female-dominated profession with 68% of female teachers in primary school, 58% in lower secondary and 52% at upper secondary.

4.1.2 Language/s

Several studies highlight the interconnected relationship between South Africa's multilingual landscape and the curriculum (Babane & Maluleke, 2020; Posel et al., 2020; Stoffelsma & Spooren, 2019; Taylor & Von Fintel, 2016). However, the extent to which languages in homes or schools contribute to student performance is not clear. Table 7 outlines the students' responses concerning their home languages, the main language of instruction at Grades 1 to 3, and the main language of instruction at high school.

Table 7 highlights the selected home languages from the most responses to the least. Of the 12 official languages, ten languages were selected as home languages by participants, reflecting the diverse linguistic landscape of the study group. Indigenous languages including isiZulu, Sepedi, isiXhosa and Setswana had the largest number of respondents. Only 5% of the learners selected English as their home language. None of the students chose Afrikaans as their home language. The high number of responses for indigenous languages as the students' home language is in line with Posel et al.'s (2020) findings that English is not commonly used language in people's everyday conversations in South Africa and that indigenous languages are the most spoken.

The Curriculum and Assessment Policy Statements (CAPS) recommends the introduction of English as a First Additional Language from Grade 1 in schools where an African language or Afrikaans is used as the Language of Learning and Teaching in the Foundation Phase. The highest number of students reported using multiple languages (including English) as their primary language of instruction in Grades 1 to 3 with 26.3%, followed by those who reported using English only as their sole language of instruction with 21.1%. To rectify the inequalities of Apartheid, the South African education policy encourages the use of all official languages and allows the School Governing Bodies to choose the language of instruction (Taylor & Von Fintel, 2016). The autonomy granted to School Governing Bodies results in significant variation in language policies across schools.

Table 7

Frequency of home languages, main languages of instruction in Grades 1 to 3 and in high school among participants

Languages	Home Language Frequency	Gr 1 to 3 Language Frequency	High School Language Frequency
isiZulu	30.1% (<i>n</i> = 40)	15.8% (<i>n</i> = 21)	5.3% (<i>n</i> = 7)
Sepedi	15.0% (<i>n</i> = 20)	9.0% (<i>n</i> = 12)	2.3% (<i>n</i> = 3)
isiXhosa	13.5% (<i>n</i> = 18)	6.0% (<i>n</i> = 8)	0
Setswana	12.8% (<i>n</i> = 17)	8.3% (<i>n</i> = 11)	3.0% (<i>n</i> = 4)
Sesotho	9.8% (<i>n</i> = 13)	5.3% (<i>n</i> = 7)	1.5% (<i>n</i> = 2)
Xitsonga	6.8% (<i>n</i> = 9)	3.0% (<i>n</i> = 4)	1.5% (<i>n</i> = 2)
siSwati	4.5% (<i>n</i> = 6)	1.5% (<i>n</i> = 2)	0
English	4.5% (<i>n</i> = 6)	21.1% (<i>n</i> = 28)	51.0% (<i>n</i> = 68)
Tshivenda	2.3% (<i>n</i> = 3)	2.3% (<i>n</i> = 3)	0
isiNdebele	0.8% (<i>n</i> = 1)	0	0
Afrikaans	0% (<i>n</i> = 0)	0.8% (<i>n</i> = 1)	0
Bi-or multi-lingual (including English)	0% (<i>n</i> = 0)	26.3% (<i>n</i> = 35)	35.3% (<i>n</i> = 47)
Bi-or multi-lingual (excluding English)	0% (<i>n</i> = 0)	0.8% (<i>n</i> = 1)	0

Some students in this study chose isiZulu, Sepedi, Tshivenda, isiXhosa, Sesotho, Setswana, siSwati and Xitsonga as their only primary languages of instruction, however, the frequency of these responses is lower compared to the home language responses. While some schools may prioritise English as the language of instruction from Grade 1, others may choose to use other official languages as their primary medium of instruction, introducing English only later. This variation can lead to unequal opportunities for learners to develop English proficiency, an important skill for academic success in later stages of their educations.

Notably, one student selected Afrikaans as their main instructional language in Grades 1 to 3, and several others included Afrikaans as one of the languages of instruction, but none of the students chose Afrikaans as their home language. Very few students chose English as their home language ($n = 6$), yet more students responded that they were taught solely in English in Grades 1 to 3 ($n = 28$) or in multiple languages including English ($n = 35$). The discrepancy in home languages and the main language of instruction in Grades 1 to 3, highlighted by Afrikaans and English, could reflect the challenges posed by a multilingual environment and the impact this could have on the students' academic outcome.

There is a rise in the number of students who selected English as their sole language of instruction in high school and several also chose multiple indigenous languages (including English) ($n = 68$ and $n = 47$, respectively), compared to the selected responses for the main language of instruction at Grades 1 to 3. Taylor and Von Fintel (2006, p. 75) argue that in South Africa, secondary school-leaving examinations are only offered in English and Afrikaans. The high number of responses for English and multiple languages (including English) corresponds with Taylor and Von Fintel's (2006) assertion that many students opt to learn English as their second language rather than Afrikaans. The increase in English as the main language of instruction may also reflect on the role of high school in equipping students with necessary skills for higher education.

Some students selected only indigenous languages as their main language of instruction in high school. As argued by Dyers and Antia (2019), providing course materials in indigenous languages gives learners a chance to better understand the content and actively participate in a language they are comfortable with. This is closely related to the Affordance Theory, which supports that when students are afforded an accommodating environment for them to learn, for example, in their home languages, they can be motivated to realise their potential. However, a study by Khanyile and Awung (2023) found that while using indigenous languages in schools

looks good on paper, its implementation is a challenge. The study found that indigenous languages have limited subject-specific terminology, teachers lack the training to teach in indigenous languages and schools lack mother-tongue resources. However, one might question whether teaching students in mainly in indigenous languages is truly beneficial, given that English remains the dominant language of ‘power’ in higher education (Dyers & Antia, 2019). This study highlights the need for investigations on the impact of teaching students solely in indigenous languages in preparing them for an English-taught university environment.

4.1.3 Reading strategies and reading habits

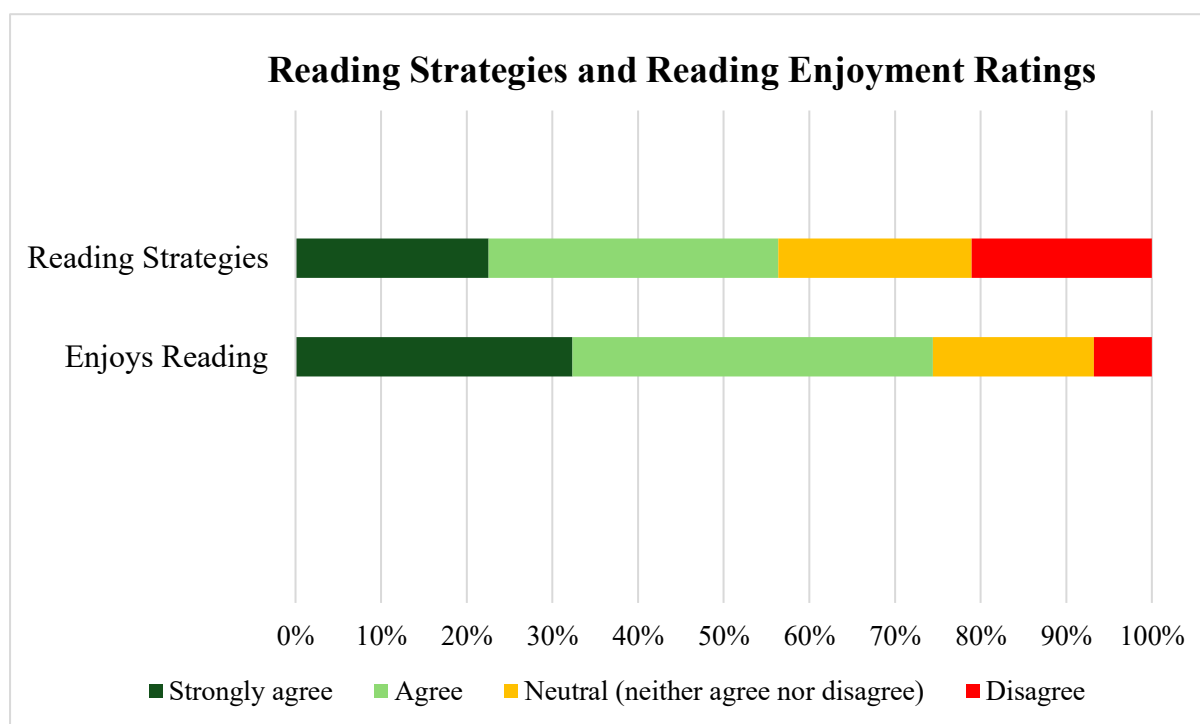
Matope and Senye-Awudi (2024) argue that teaching reading strategies in high school is necessary to help students read and comprehend text independently, which will help them achieve academic success. A significant number of students strongly agreed or agreed (56%) that they were taught reading strategies in high school. There are students (21%) who indicated that they were not taught reading strategies in high school; this could suggest a gap in curriculum or teaching approaches, potentially impacting the students’ reading skills at tertiary level. The ‘Neutral (neither agree nor disagree)’ response suggests a need for further interpretation; this suggests that students may be unsure if they were taught reading strategies, are unfamiliar with what reading strategies are, or feel the strategies taught were inadequate, leading to the neutral response.

A total of 74.4% of students strongly agreed or agreed that they enjoy reading, indicating a positive attitude towards reading. Findings in Figure 1 show that the number of students who responded that they enjoy reading ($n = 99$) is higher compared to the number of students who responded that they were taught reading strategies in high school ($n = 75$); this suggests that many students could have developed an enjoyment of reading that does not stem from formal reading strategies instruction in high school. This warrants a need for further investigation into other potential reasons that could influence students’ enjoyment of reading. Anwar and Sailuddin (2022) argue that students who do not habitually read for pleasure often face challenges when reading other types of texts, like academic texts.

Figure 1 shows that 6.8% students indicated that they did not enjoy reading and this can stem from a variety of aspects, such as lack of reading resources, difficulties reading and understanding reading materials, or distractions from different digital media content.

Figure 1

Participants' responses to the statements: 'I was taught reading strategies in high school' and 'I enjoy reading'



The questionnaire asked students to indicate how well they think they read and write in 11 official languages (SASL was not recognised as an official language at the time the data was gathered); the current study only focuses on English because most South African universities teach in English. As shown in Table 8, most of the participants responded that they read and write very well in English; this indicates a strong level of confidence amongst students regarding their English language proficiency, which could indicate an effective foundational and high school instruction in English. As Kruger and Dunning (1999) argue in the Dunning-Kruger effect theory, people who are unskilled in social and intellectual domains tend to hold overly favourable views of their competencies, which could explain why most students in the study reported that they read and write well, despite other indicators suggesting challenges with academic literacy. While this may be the case, it is equally important to acknowledge that the questionnaire asks students how well they read and write, which does not accurately indicate whether they are overestimating their reading or writing abilities. Therefore, this warrants further investigation to understand why students believe their competencies are high, whereas the results suggest otherwise.

The number of students who responded that they read and write very well in English is almost on par with the number of students who reported that they strongly agree and agree that they enjoy reading. Dedicated reading motivates students to develop strategies to encourage further reading (Spjeldnæs & Karlsen, 2024), which could result in student confidence in their reading and writing abilities. This could be an indication of the connection between reading for pleasure as well as reading and writing proficiency.

Table 8

Frequency of how well students read and write in English

Read and Write Proficiency [English]	Frequency
Very well	75.9% (<i>n</i> = 101)
Comfortably	20.3% (<i>n</i> = 27)
Only a little bit	1.5% (<i>n</i> = 2)
I don't read and write in this language	2.3% (<i>n</i> = 3)

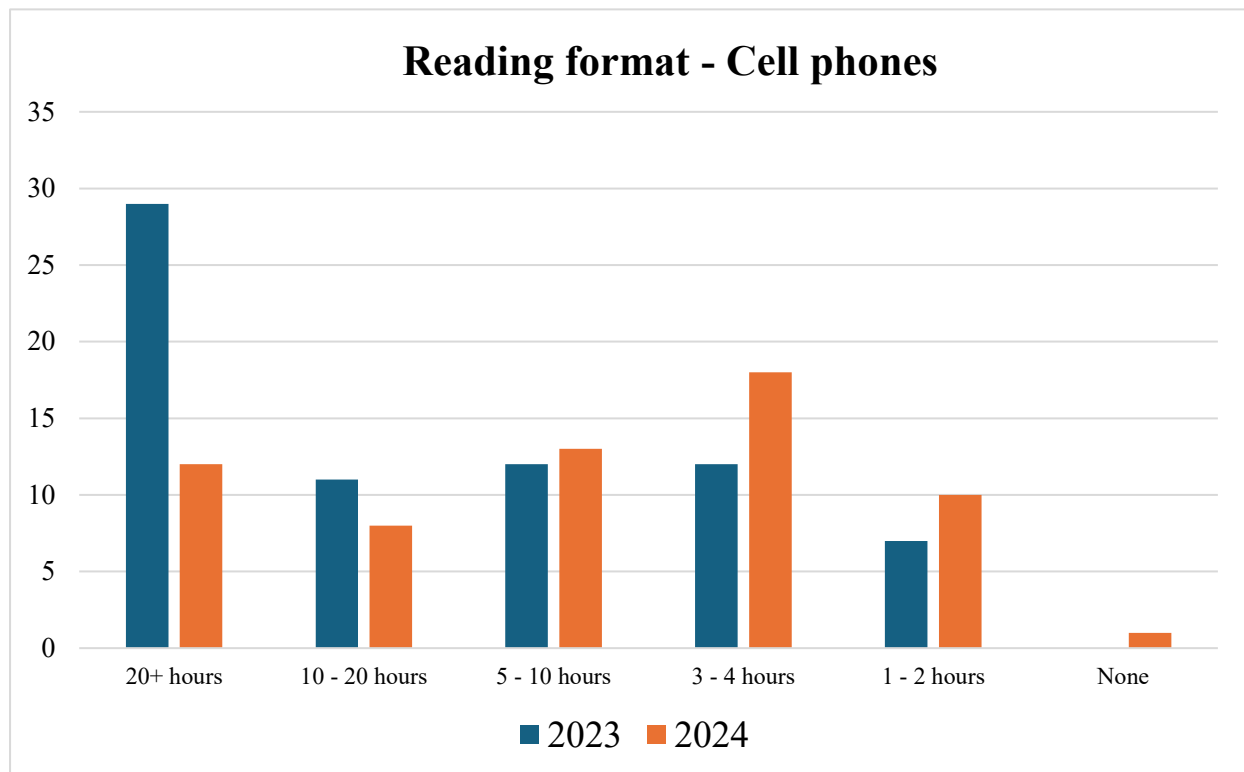
As Bharuthram (2017) highlights, forming a habit of regular reading can be helpful for students, as it helps them to acquire more vocabulary, and enhance their reading comprehension which ultimately improves their reading and writing abilities. There are three students who responded that they do not read or write in English, raising critical concerns about their ability to succeed in an English-medium academic environment.

4.1.4 Reading formats

The varied patterns of dependence on cell phones, laptops and hardcopies shown in Figures 2, 3 and 4 demonstrate how affordances show how students are afforded different reading devices that influence their reading patterns and habits (Spjeldnæs & Karlsen, 2024). Figure 2 shows that in 2023, there is a notably high number of responses from students (41%) who spend more than 20 hours per week reading on their cell phones. This suggests that a significant portion of the students rely heavily on their cell phones for reading. The response frequencies decrease as the number of hours decrease, which could be an indication of accessibility and convenience of mobile devices and how much students rely on them for their reading activities.

Figure 2

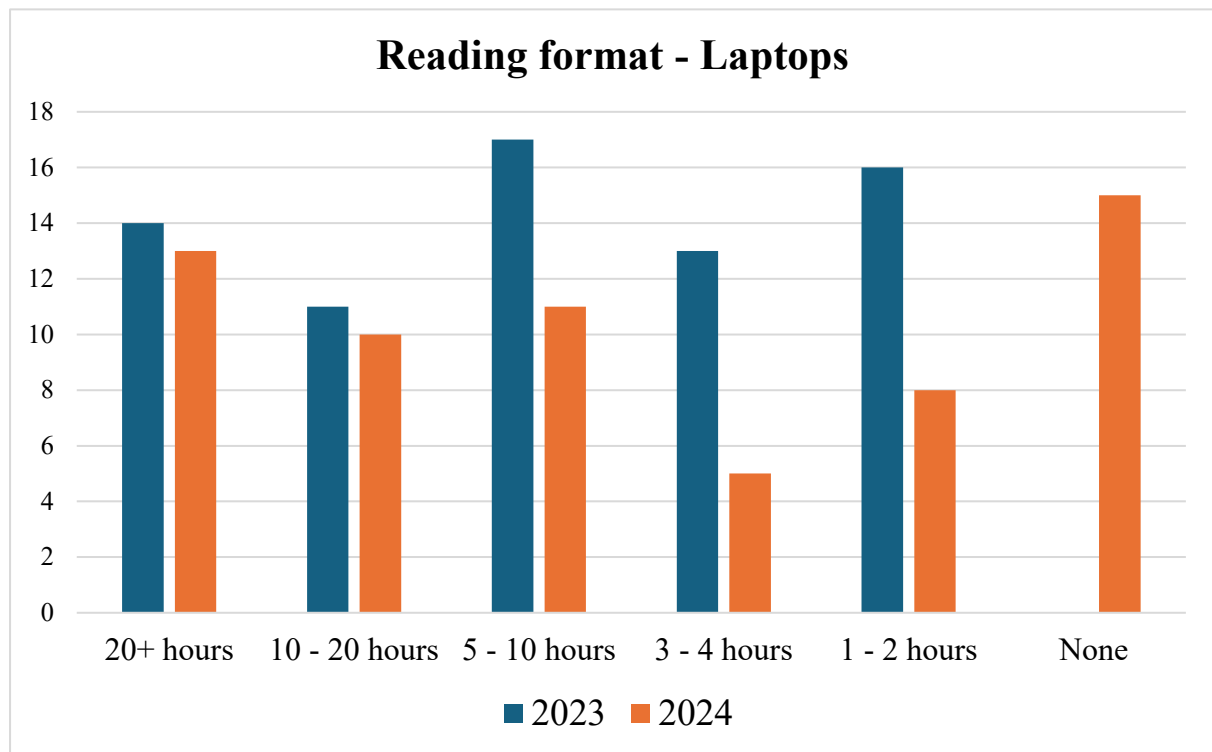
Frequency of time per week spent reading on cell phones



Responses for time students spend reading on their laptops are more spread out across categories from more than 20 hours to none, indicating a wide variation in reading behaviours on laptops (see Figure 3). In 2024, the highest number of students responded that they do not spend any time reading on laptops (24.2%), which could be a further indication that more time is spent reading on other formats. With over 60% of South African children living in poverty (Nortje, 2017), the lack of time spent reading on laptops could be an indication that some students do not have access to laptops as a result of financial constraints. The researcher also noted that the question specifically asks about reading; students may be using laptops for other academic tasks, such as writing assignments, watching tutorials or online discussions.

Figure 3

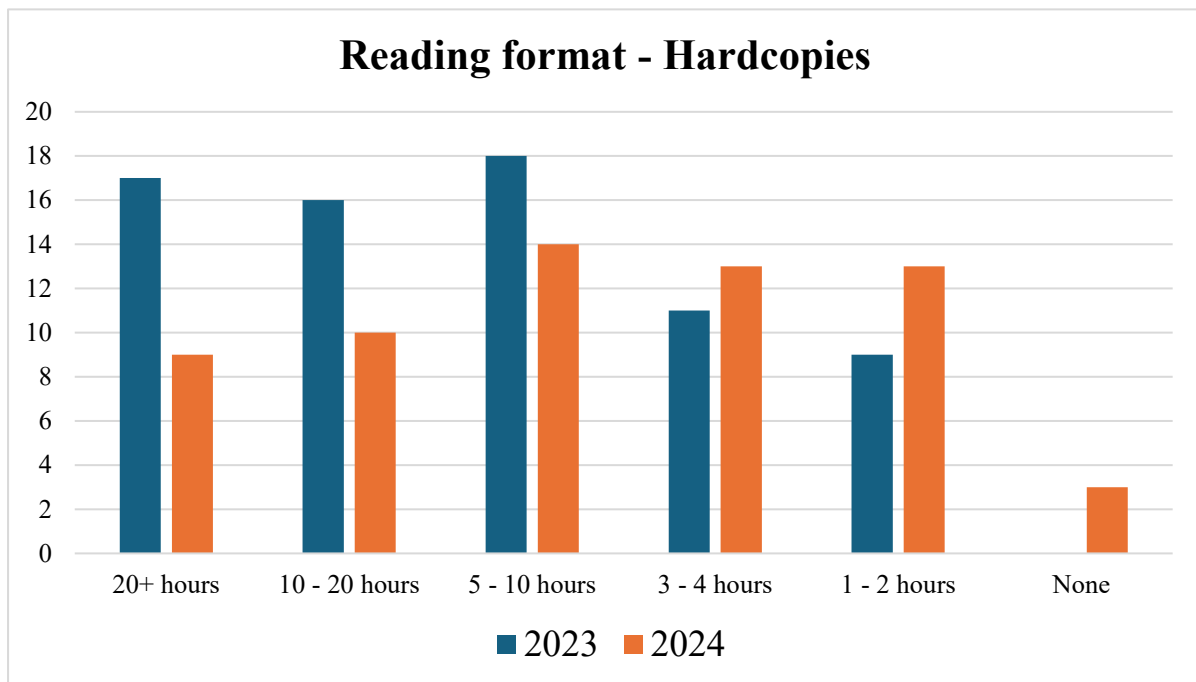
Frequency of time per week spent reading on laptops



The responses for time spent reading hardcopies are relatively evenly spread across the categories, as shown in Figure 4. This highlights that while technological affordances like cell phone and laptops are becoming prominent in education, the traditional hardcopy format cannot be easily replaced. The type of text influences the choice of reading format, for example, digital reading favours lighter texts while complex texts are easier to navigate in hardcopy format (Spjeldnæs & Karlsen, 2024). The lack of a clear trend on how many hours students spend on reading hardcopies could be a further indication of the availability of different reading formats; therefore, the time spent on different formats creates a diverse pattern of reading habits.

Figure 4

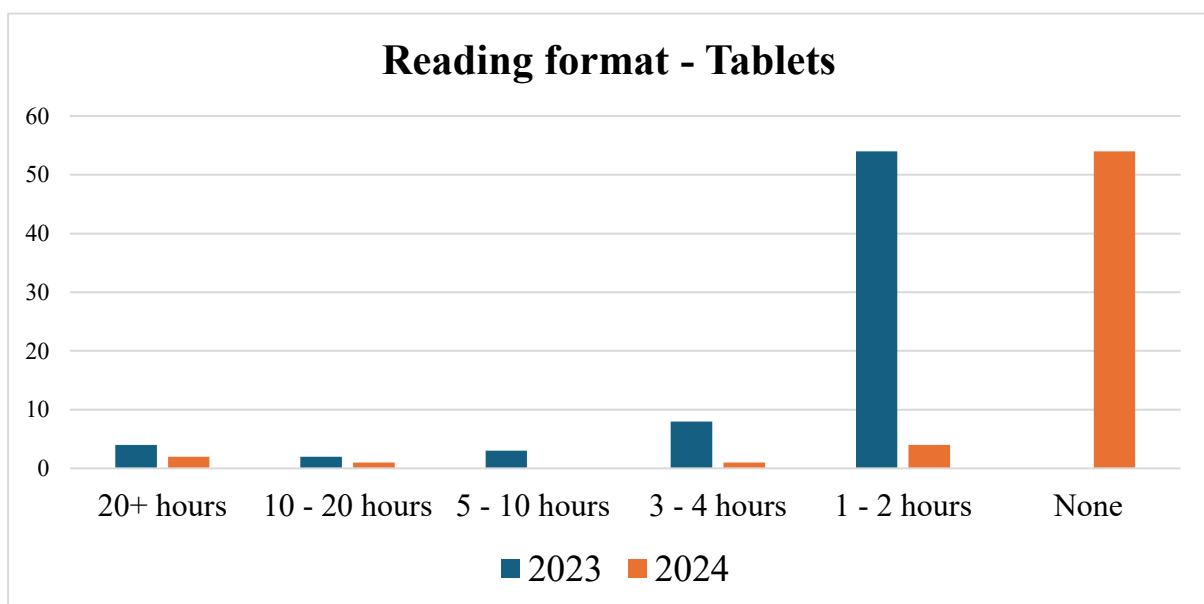
Frequency of time per week spent reading on hardcopies



The most notable trend between the two cohorts is for time spent reading on tablets, as shown in Figure 5, with a high number of responses for the '1 – 2 hours per week' concentrated in the 2023 cohort and the 'None' responses in the 2024 cohort.

Figure 5

Frequency of time per week spent reading on tablets



Though both responses show that students spend little to no time reading on tablets, one can conclude that the shift from ‘1 – 2 hours per week’ in 2023 to ‘None’ in 2024 (as indicated in Section 3.4) could be a reflection that the 2024 options allowed students to accurately capture their responses whereas in 2023 students chose a default answer that is close to the little or no time spent reading on tablets.

4.1.5 Word level scores

To demonstrate mastery of vocabulary at all levels, a score of 15 or 16 out of 18, which is equivalent to approximately 85% to 90%, is considered the threshold (Laufer & Nation, 1999). Table 9 shows the minimum score, first quantile, median, mean, third quantile, maximum score and the number of students who got 85% and above.

Table 9

Vocabulary word level results

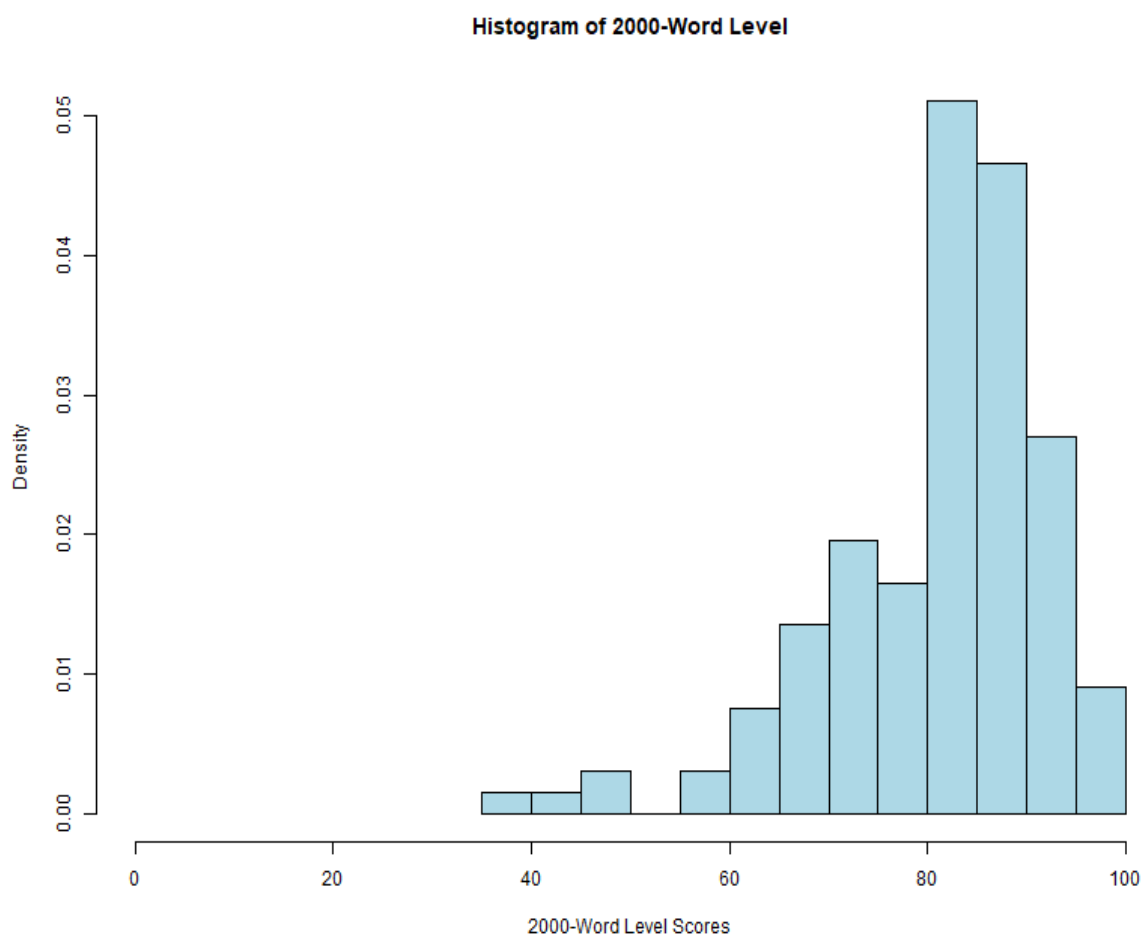
Vocabulary Word Levels	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	85% and above
2000-word level	38.89%	77.78%	83.33%	81.31%	88.89%	100.00%	55 (41.4%)
3000-word level	25.00%	58.33%	72.22%	69.30%	83.33%	94.44%	25 (18.8%)
5000-word level	11.11%	38.89%	50.00%	49.39%	61.11%	83.33%	0
UWL	19.44%	52.78%	61.11%	58.73%	66.67%	86.11%	1 (0.8%)

The scores for the 2000-word level range from a minimum of 38.89% to a maximum of 100%. A notable 41.4% of students scored above the required 85% mastery mark, which corresponds with Laufer et al.’s (2004) assertion that the 2000-word level is the easiest level. The mean score of 81.31 % falls just below the mastery threshold, indicating that many students ($n=78$) did not meet the threshold. The distribution for the 2000-word level score is left-skewed (see

Figure 6), which means that there are more scores concentrated towards the higher values and few scores that are on the lower end. This is an indication that students performed relatively well on the 2000-word level, though many of the students did not achieve the mastery threshold. This further implies that more than 50% of the students did not know most of the high frequent words in 2000-word level which implies that students are less likely perform well in reading comprehension (Stæhr, 2008).

Figure 6

Histogram of 2000-word level results

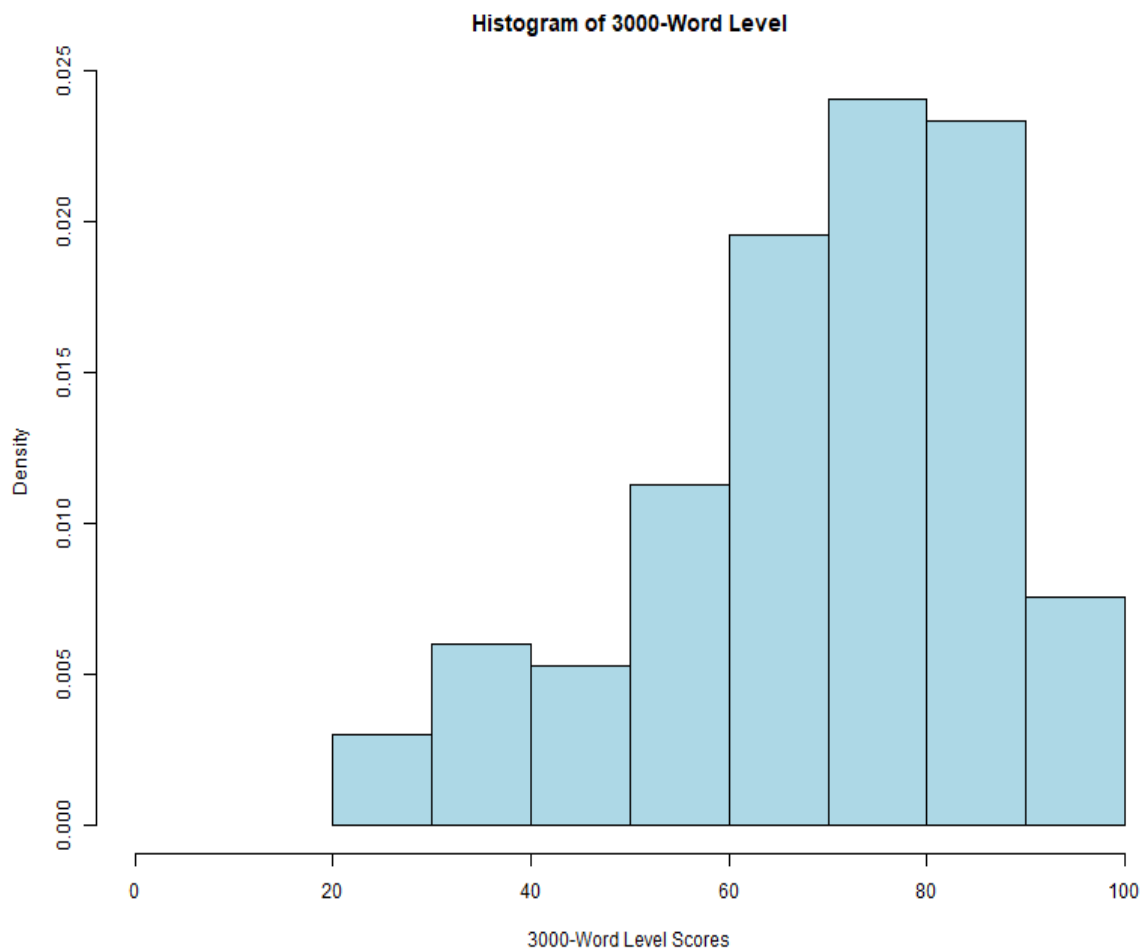


The scores for the 3000-word level are lower than the 2000-word level. Only 18.8% of the students managed to achieve the proficiency level, compared to the 41.4% in 2000-word level. Though the 2000- and 3000-word levels represent the foundational vocabulary level (Scheepers, 2016), the difference in the scores highlights the progression from the 2000-word level which is the easiest level to the 3000-word level which is a slightly more challenging level in the development of vocabulary proficiency. The mean of 63.30% for the 3000-word

level is relatively lower than the mastery level and over 80% of participants fell short of the 85% mastery threshold. As shown in Figure 7 shows a left-skewed distribution suggesting that a significant portion of the students performed fairly well; however, reaching the threshold level remains a challenge for most of the participants.

Figure 7

Histogram of 3000-word level results

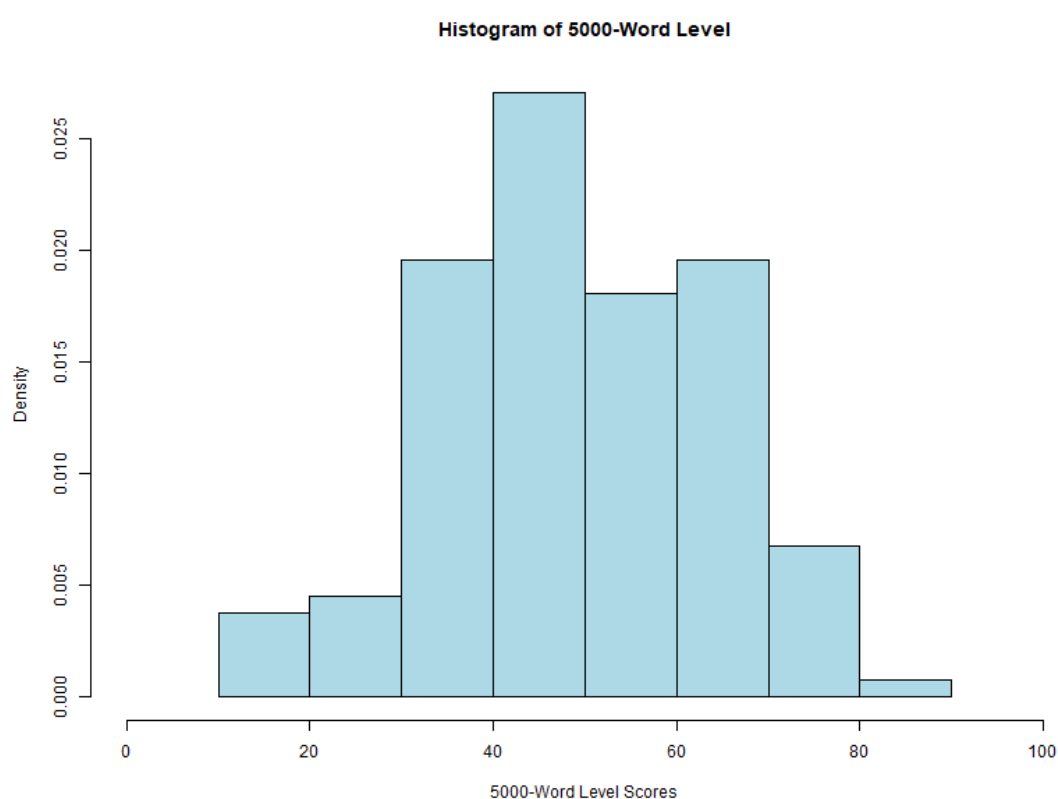


The results for the 5000-word level show a distribution of student scores ranging from a low score of 11.11% to a high score of 83.33%. None of the students managed to achieve the 85% mastery level. Laufer et al. (2004) argue that the 3000-word level is an important stage before progressing to the 5000-word level, as this is where students bridge the gap between assisted reading and progress to unassisted comprehension (Nation & Anthony, 2013). The significant difference between the 3000- and 5000-word level scores suggests a gap in students' ability to transition from structured learning to independent learning and indicates a need for further research. This also highlights the difference between BICS and CALP, showing the difference

between conversational vocabulary and taught academic vocabulary. The histogram of the 5000-word level scores (see Figure 8) has a normal distribution, which implies that the highest frequency of values occurs near the mean, suggesting that there is little to no skewness. The distribution further highlights the challenge of mastering the 5000-word level for all participants. This could be an indirect reflection of students' reading habits, as those who read less are unlikely to encounter and acquire mid-frequency words through exposure, whether in spoken or written language.

Figure 8

Histogram of 5000-word level results



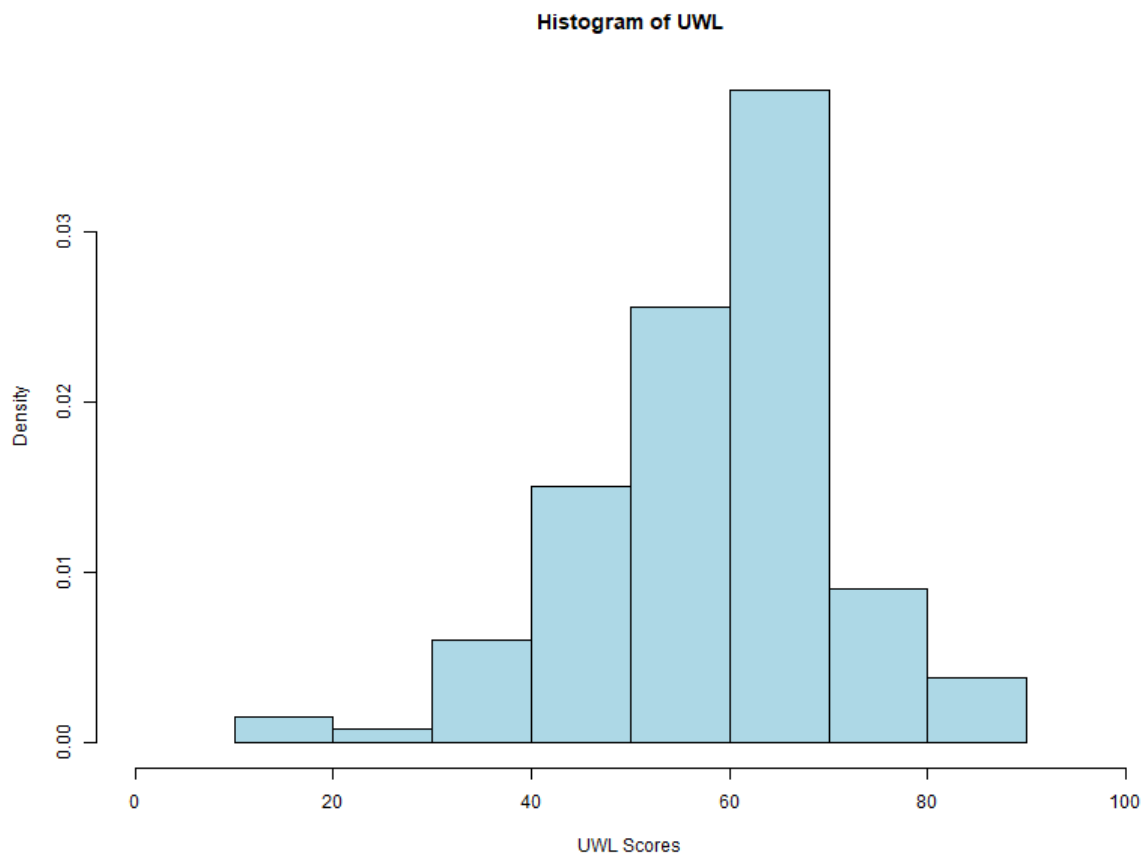
The UWL is considered the benchmark for academic vocabulary proficiency for this study because it serves as an essential measure of students' academic abilities at tertiary level (Gholaminejad, 2020). The scores for the UWL range from a low of 19.44% to a high of 86.11%, with only one student achieving mastery proficiency of 85%. Laufer et al. (2004) argue that knowledge of the 5000-word level would precede mastery of the UWL; however, the results in Table 9 seem to indicate a slightly reversed trend, where students seem to perform slightly better in the UWL than in the 5000-word level. The DHET (2021) argues that while academic vocabulary is mainly important for higher education, it is important for students to

access academic vocabulary in their high school senior phase. This could be an indication that learners in South Africa, with its unique second language context, are exposed to more academic vocabulary in high school, due to curricula that encourage this, than would be the case for the typical English as a Foreign Language speaker.

The distribution is slightly left-skewed as shown in Figure 9, indicating a concentration of scores near the higher end, with a few lower values pulling the tail to the left. While this pattern would usually suggest a relatively balanced and moderately strong performance, in this case, the high peak around the 60% mark is significantly below the 85% mastery threshold.

Figure 9

Histogram of UWL results



The UWL consists of specialised vocabulary of university texts (Nation, 1983); therefore, a lack of mastery of the UWL would imply that students would struggle with comprehending their academic texts.

4.2 Bivariate analysis results

The second objective of the study was to explore the associations between academic vocabulary proficiency and each selected predictor among first-year university students. The results of the bivariate analysis revealed four variables that have statistically significant relationships with the UWL score; these variables include students who disagreed that they were taught reading strategies students in high school, 2000-, 3000- and 5000-word levels. Other predictors, namely age, gender, home language, how well students read and write in English, their main languages of instruction at Grades 1 to 3 and in high school, how much they enjoy reading were not statistically significant.

The regression results in Table 10 show a statistically significant association between the UWL score and students who disagreed that they were taught reading strategies in high school. Students who disagreed are predicted to score 5.79 points lower on the UWL test than those who agreed (predicted score = 55%). The intercept suggests that on average students who agreed that they were taught reading strategies in high school are predicted to score about 60.9% on the UWL test. The predicted score of 60.9% for students who agreed that they were taught reading strategies is much lower than 85% mastery threshold.

Table 10
Simple regression model: UWL ~ Reading strategies

	Dependent variable: UWL
reading_strategiesNeutral (neither agree nor disagree)	-4.000 (2.621)
reading_strategiesDisagree	-5.793** (2.687)
Constant	60.852*** (1.401)
Observations	133
R ²	0.041
Adjusted R ²	0.026
Residual Std. Error	12.132 (df = 130)
F Statistic	2.788* (df = 2; 130)

Note: *p<0.1; **p<0.05; ***p<0.01

Sebolai's study (2022) that highlights a knowledge gap that exists between high school and tertiary education, where learners are leaving high school without the necessary skills needed to succeed academically and negative association between UWL the high number of students who have indicated that they were not taught reading strategies in high school highlights could be a further indication of this gap.

The intercept for the 2000-word level and UWL scores suggests that when the 2000-word level is zero, students are predicted to score 11.7% in the UWL. The positive, significant, coefficient, as shown in Table 11, further predicts that a one-unit increase in the 2000-word level score can be associated with a 0.58 unit increase in the UWL score. The positive association indicates the value of 2000-word level as an important predictor of academic vocabulary proficiency.

Table 11

Simple regression model: UWL ~ 2000-word level

	Dependent variable: ----- UWL
wordlevel2000	0.579*** (0.081)
Constant	11.689* (6.660)
Observations	133
R2	0.280
Adjusted R2	0.274
Residual Std. Error	10.476 (df = 131)
F Statistic	50.842*** (df = 1; 131)

Note: *p<0.1; **p<0.05; ***p<0.01

The 2000-word level represents BICS (Webb et al., 2017), which is acquired during the foundational stages before formal schooling, highlighting the importance of socially acquired vocabulary knowledge in a student's overall academic vocabulary proficiency.

It can be seen in Table 12 that there is a positive, significant association between the 3000-word level and UWL, this predicts that where there is an increase of one-unit in the 3000-word level it can be associated with a moderate 0.41 increase in the UWL score. Students are usually introduced to words in the 3000-word level at the foundational stage of school and this is when

they start to develop CALP (Scheepers, 2016, p. 56), thus highlighting the importance of a strong foundational phase that will carry on to academic vocabulary proficiency at higher education.

Table 12

Simple regression model: UWL ~ 3000-word level

Dependent variable:	
UWL	
wordlevel3000	0.409*** (0.052)
Constant	30.411*** (3.681)
Observations	133
R2	0.324
Adjusted R2	0.319
Residual Std. Error	10.149 (df = 131)
F Statistic	62.763*** (df = 1; 131)

Note: *p<0.1; **p<0.05; ***p<0.01

The simple regression analysis also confirms a positive, significant association between the 5000-word level and UWL. Initially, the diagnostic tests for the simple regression for the UWL score and the 5000-word level scores show one outlier at point 75 and this was identified as row 69 (see results in Appendix B). This value displays as an outlier because this is highest score achieved in the 5000-word level category. The substantial difference between the highest score (83.33%) and the second-highest score (77.78%) may explain why point 75 was identified as an outlier. The researcher decided to run another regression without point 75 and the revised results show a higher coefficient than in the previous regression that had the influential point.

As shown in Table 13, the model predicts that for every one-unit increase in the 5000-word level, the UWL score is expected to increase by an average of 0.53 units. The 5000-word level represents an advanced level of CALP where students transition from assisted reading to independent reading which is a requirement for reading comprehension (Nation & Anthony, 2013). This association highlights the importance of high school in providing students with the necessary skills that will equip them for higher education.

Table 13*Simple regression model: UWL ~ 5000-word level*

	Dependent variable: ----- UWL
wordlevel5000	0.530*** (0.058)
Constant	32.845*** (2.966)
Observations	132
R2	0.392
Adjusted R2	0.387
Residual Std. Error	9.537 (df = 130)
F Statistic	83.737*** (df = 1; 130)

Note:

*p<0.1; **p<0.05; ***p<0.01

4.3 Multiple linear regression results

The third objective of the study was to examine the extent of the association between academic vocabulary proficiency and the statistically significant predictors identified in the simple regressions. The results for the multiple linear regression revealed that the four independent variables that were significant in the simple regressions (see Section 4.2) are still significant when holding other variables constant. The results are as follows for the statistically significant predictors are shown in Table 14.

There is statistically significant evidence that the UWL scores for students who disagree that they were taught reading strategies in high school are, on average, 5.07 points lower than those who agree, after controlling for other statistically relevant predictors included in the model. This indicates that students who reported not receiving instruction in reading strategies during high school tend to have significantly lower academic vocabulary scores. The difference between the coefficient of -5.8 in the simple regression and -5.1 in the multiple regression is small, which suggests that there is a meaningful association between UWL and the lack of reading strategies instruction in high school, even when accounting for potential confounding variables. The negative association confirmed by both the simple and multiple regressions,

suggests that not being taught reading strategies has a negative impact on academic vocabulary proficiency, which could have an overall negative impact on academic reading.

Table 14

Multiple Regression Model: UWL ~ Reading strategies, 2000-, 3000- and 5000-word level

	Dependent variable: UWL
reading_strategiesNeutral (neither agree nor disagree)	-3.714* (1.966)
reading_strategiesDisagree	-5.067** (2.001)
wordlevel2000	0.201** (0.089)
wordlevel3000	0.169*** (0.064)
wordlevel5000	0.312*** (0.073)
Constant	17.411*** (5.859)
Observations	132
R2	0.507
Adjusted R2	0.487
Residual Std. Error	8.724 (df = 126)
F Statistic	25.887*** (df = 5; 126)

Note: *p<0.1; **p<0.05; ***p<0.01

The results for the 2000-, 3000-, and 5000-word levels were all statistically significant, even after controlling for other predictors in the model. This indicates that each of these word levels contributes uniquely to the prediction of academic vocabulary proficiency. However, compared to the initial simple regression analyses results, which suggested moderate positive associations between each word level and UWL, the associations observed in the multiple regression analysis were notably weaker, indicating weak positive associations. This decline in the strength of the associations suggests that the possibility of intersections between the word

levels, hence highlighting that it is necessary for students to progress from one word level to another until they reach the UWL phase. Mastery at one word level supports and overlaps with the development of the next.

The diagnostic tests show a point of influence around the 68th point that is at -4 on the studentised residuals plot, and visible on the Cook's Distance plot (see results in Appendix C). This outlier is the same point that was picked up on the UWL and 5000-word level simple regression (row 69). Another regression was run without the influential point and the revised results remained statistically significant with a slight difference in the estimates.

The adjusted R-squared of 0.49, as shown in Table 14 means that the model accounts for nearly 49% of variance in the UWL, which suggests a moderate fit of the data to the model. This level of explained variance confirms that the predictors included in the model contribute meaningfully to explaining academic vocabulary proficiency. The remaining unexplained variance highlights the possibility of additional factors not captured in the current model, such as reading habits, foundation learning, or language/s in academic vocabulary proficiency. Nonetheless, the moderate fit highlights that the statistically significant predictors play an important role in shaping academic vocabulary outcomes.

4.4 Chapter summary

This chapter presents and interprets the findings of the study, focusing on the statistical analyses conducted to address the research objectives. The chapter is divided into sections detailing the univariate, bivariate, and multiple linear regression analyses. While the univariate analysis presents the multifaceted learning environments that students are exposed to, the bivariate and multiple regression analyses reveal that the lack of reading strategies instruction and knowledge of the 2000-, 3000- and 5000-word levels are associated with UWL. The analyses confirm that some predictors meaningfully contribute to explaining academic vocabulary proficiency, although the moderate explanatory power suggests that there may be other factors that also play a significant role. This underscores the need for further investigation into unexplained variance and underlying factors impacting student performance.

Chapter 5: Conclusions and Recommendations

5.1 Introduction

This study's objectives were to identify the predictors related to academic vocabulary proficiency, explore associations and the extent of relations between academic vocabulary proficiency and the selected predictors among first-year university students. This chapter summarises the findings of the study and concludes the research. The chapter also provides recommendations for the data collectors and future research.

5.2 Summary of main findings

This study set out to explore these predictors and their associations with academic vocabulary proficiency among first-year university BEd students. The findings highlight complex and multifaceted learning environments in South African schools, demonstrating that students' academic vocabulary development is shaped by multiple interconnected factors within these environments.

The simple and multiple regression results for the UWL and students who disagreed that they were taught reading strategies in high school align with literature highlighting potential gaps between the skills taught in high school and those required for academic success. There is a need to explore whether students are familiar with reading strategies and understand how to apply them effectively. This highlights the importance of seeking specific pedagogical affordances and some unrealised factors in the learning environments (Moll et al., 2022), for example, students' neutral response when asked if they were taught reading strategies in high school could be an indication of unrealised factors. The negative results from the regression analyses suggest that the lack of reading strategy instruction at the high school level may have had a negative impact on students' academic vocabulary proficiency.

The results in this study show a trend where students' VLT scores decrease as word levels increase. Drawing on Cummin's (2008) theory, this also highlights the difference between BICS and CALP. This suggests that students are more familiar with the high-frequency words usually found in the 2000- and 3000-word levels, while mid- or low-frequency words in the 5000-word level and UWL present challenges for students. Laufer and Nation (2001) argue that vocabulary knowledge encourages fluency which is expected in everyday conversations and reading as one's vocabulary size increases. There is a need to understand which vocabulary

increases, conversational or academic vocabulary, and how this contributes to academic reading and academic success, at large. The gradual decline in knowledge of academic words highlights the importance of targeted vocabulary teaching to support students' academic reading skills.

All word levels are statistically significant in the simple regression models indicating that each word level (2000, 3000, and 5000) has a measurable and meaningful relationship with students' UWL scores. This further suggests that vocabulary knowledge at each of these levels contributes significantly to academic vocabulary proficiency. In other words, the more words students know at each frequency level, the better they tend to perform in academic vocabulary tasks. This finding highlights the importance of vocabulary development across different levels of word frequency, reinforcing the role of both BICS and CALP in supporting academic success. The moderate associations could be an indication that there are other factors beyond vocabulary knowledge that influence reading proficiency, as suggested by the Affordance Theory. Ngwenya (2010) argues that general language competence is not a dependable predictor of a student's academic success, so there is a need to look at other factors like the influence of foundational learning in tertiary education.

The decline in the associations between the UWL and the 2000-, 3000-, and 5000-word levels in the multiple linear regression suggests a nuanced relationship between these word levels, which is essential for students to progress from one level to another and influences their academic vocabulary proficiency. The results highlight the importance of students achieving mastery across all word level thresholds, as each level builds upon the previous one, contributing to the overall development of academic vocabulary. This interconnected progression is crucial for understanding how vocabulary acquisition occurs and how it impacts students' academic success. The reduced association may also suggest that language proficiency alone does not fully explain academic vocabulary proficiency and as Cummins (1979) further suggests there is a need to explore socio-cultural variables that may influence academic outcomes rather than just linguistic factors.

Importantly, while some predictors like age, gender, languages, reading habits and how well students read and write in English were not found to be statistically significant in the simple regression models, this does not necessarily indicate the absence of meaningful associations with UWL. Instead, it highlights the limitations of the current data and methods in fully capturing these relationships and points to the need for further research using more nuanced

models that account for socio-cultural, linguistic, and contextual factors. Ultimately, this study contributes to the growing body of research calling for a holistic understanding of vocabulary development within South Africa's educational landscape. A descriptive understanding of predictors that relate to academic vocabulary proficiency may influence further research that can investigate the gaps identified and questions asked.

5.3 Limitations of the study

This study strived for a thorough investigation of the data to provide rigorous results; however, it is important to highlight some inherent limitations of the study. Firstly, the data were only collected from a single South African university, and only from two cohorts of Bachelor of Education (BEd) students in this university. Hence, the validity of the findings will be delimited to only Wits and to the 2023 and 2024 first-year cohorts that took part in the study. Secondly, academic reading has many components, but the study is delimited to academic vocabulary.

The chosen instruments also have some inherent limitations. Webb and Sasao (2013) argue that the VLT's limitation is its inability to capture other word levels such as the 1000- and 4000-word levels. Words in the 1000-word level are high frequency level words used more often in everyday conversations and have a greater impact on second language learners' comprehension and production; therefore, failure to capture this level hinders a more accurate placement of learners based on their vocabulary skills (Webb & Sasao, 2013). Additionally, this was a voluntary questionnaire that students completed at home, in their own time, as part of their course, which may explain why the study did not have a 100% response rate. Moreover, there is no guarantee that the students themselves completed the questionnaire, which presents a potential limitation in terms of data reliability and participant authenticity. Furthermore, a limitation of the VLT version used for this study is that students often provide various forms of the word. To address this, students were awarded half marks for partially correct answers, indicating they recognised the word, and full marks for completely correct answers, demonstrating both recognition and proper usage of the word. Furthermore, the RHQ is based on students' own perception of their reading abilities, which is often unrealistically high (Weder, 2022), and can result in biased responses.

Finally, while a multiple linear regression acknowledges other factors that can influence a relationship, it is important to note that correlation or association does not mean causation (Bertani et al., 2018). While exploratory research can yield some results that can reveal a new

phenomenon, Swaraj (2019) argues that one of its disadvantages is that the answers provided are usually inconclusive. Establishing causation requires a more rigorous investigation which is beyond the scope of this study. Notwithstanding the abovementioned limitations, the research can inform further research on the factors associated with academic vocabulary levels and influence interventions to improve vocabulary as an important aspect of academic reading and writing.

5.4 Recommendations

The following recommendations are proposed for data collectors and researchers interested in further research, with the aim of addressing the identified gaps and strengthening future studies. Currently, only 11 official languages are listed in the RHQ; however, SASL has recently been added as a 12th South African official language; therefore, this should be formally added to the VLT and RHQ list of recognised languages. This addition would ensure that all official languages of the country are acknowledged, promoting greater equity and representation in education. The findings show that 26.3% of students reported using multiple languages (including English) as their primary language of instruction in Grades 1 to 3, followed by 21.1% who used English only. Although the datasets do not include details on school quintiles, these patterns may reflect trends in the schooling system, where Quintile 1 to 3 schools typically use an African language as the language of instruction in the Foundation Phase, and Quintiles 4 and 5 schools are more likely to use English. To strengthen future analyses, data collectors should record the quintile classification of each learner's primary school. This additional data would allow for more nuanced interpretations of language exposure and its implications for academic development.

The high proportion of 'Neutral' responses to whether students were taught reading strategies in high school, highlights a potential area for further studies because this could be an indication that students felt that the strategies they were taught were inadequate or they are unfamiliar with what reading strategies are. Further research is needed to better understand how reading strategies are introduced in high school and whether students clearly understand how to apply these strategies in academic settings. It is also important to consider school quantiles and their potential associations with UWL. School quantiles typically provide additional information about the school in terms of resources, teacher-training, and learning environment, which could impact vocabulary development. Several studies have highlighted that university support programmes are often too generic, with some placing greater emphasis on writing. It is

therefore recommended that universities also prioritise vocabulary development as a key component of academic support.

Furthermore, the RQH asks students how well they read and write. The data collectors need to separate the reading and writing competencies because though connected, these are distinct skills and by separating these competencies, data collectors can better understand the unique factors that influence students' performance, which may lead to more targeted interventions and educational strategies. A correlation or lack thereof between reading or writing skills and their vocabulary proficiency can be used to either support students' self-reported competencies or to identify possible overestimations of their abilities, as suggested by the Dunning-Kruger effect.

Terasawa (2024) noted that in non-English-speaking countries, English proficiency is widely linked to socio-economic status, closely linked to employment status and household income levels. In a country like South Africa where English is an additional language for most of the population, while also being the primary language of education, government, and commerce, this reality is even more pronounced. While there has been a lot of research on the impact of language on education, there is a need to understand the nuanced relationship between English and socio-economic status. In South Africa's multilingual context, reading challenges are often dismissed as mere language barriers, yet Owen-Smith (2010) suggests this oversimplifies a complex issue.

The researcher also recommends testing associations between other word levels and the various predictors to gain a more comprehensive understanding of the relationships at play. There is a possibility that certain word levels, such as the 2000- or 3000-word levels, may be more strongly associated with other predictors like the main language of instruction at Grades 1 to 3, that were not fully explored in the current analysis due to the scope of this research report. Additionally, the students in this study seem to perform better in the UWL than in the 5000-word level, while the literature (Laufer et al., 2004) suggested that the 5000-word level would typically precede the UWL stage. By investigating these potential associations, additional explanatory factors could be uncovered that may help clarify the observed reduction in estimates between the UWL and the word levels, and that may shed more light on South African university students' vocabulary progression.

5.5 Conclusion

Ultimately, this study contributes to the growing body of research calling for holistic understanding of vocabulary development within South Africa's educational landscape. Recognising that by providing a descriptive understanding of predictors that are associated with academic vocabulary proficiency, further research can be conducted to investigate the gaps identified correlations and questions raised. The reading crisis in South Africa remains a serious and persistent challenge that affects the entire education system, from the foundational phase through to higher education. While there is extensive research on the role of vocabulary in reading comprehension, little attention has been given to understanding the specific factors that shape vocabulary acquisition and proficiency. Identifying these predictors is essential for developing targeted interventions that can enhance students' vocabulary development, their overall academic literacy and academic success.

Reference List

- Almeida, A., Loy, A., & Hofmann, H. (2018). *Ggplot2* compatible quantile-quantile plots in R. *The R Journal*, 10(2), 248–261. <https://doi.org/10.32614/RJ-2018-051>
- Andrianatos, K. (2019). Barriers to reading in higher education: Rethinking reading support. *Reading & Writing*, 10(1), 1–9. <https://doi.org/10.4102/rw.v10i1.241>
- Anwar, I. W., & Sailuddin, S. P. (2022). Academic reading difficulties in higher education. *Journal of Languages and Language Teaching*, 10(2), 309–314. <https://doi.org/10.33394/jollt.v10i2.4849>
- Aukerman, M. (2007). A culpable CALP: Rethinking the conversational/academic language proficiency distinction in early literacy instruction. *The Reading Teacher*, 60(7), 626–635. <https://doi.org/10.1598/RT.60.7.3>
- Babane, M. T., & Maluleke, M. J. (2020). The role of multilingualism in an academic milieu at a selected South African university. *South African Journal of African Languages*, 40(1), 47–52. <https://doi.org/10.1080/02572117.2020.1733823>
- Babbie, E., R. (2010). *The Practice of Social Research* (12th ed.). Wadsworth.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Baumann, J. F., & Graves, M. F. (2010). What is academic vocabulary? *Journal of Adolescent & Adult Literacy*, 54(1), 4–12.
- Bayazidi, A., & Saeb, F. (2017). Assessing reliability of two versions of vocabulary levels tests in Iranian context. *Advances in Language and Literary Studies*, 8(1), 30–43. <https://doi.org/10.7575/aiac.all.v.8n.1p.30>
- Bertani, A., Di Paola, G., Russo, E., & Tuzzolino, F. (2018). How to describe bivariate data. *Journal of Thoracic Disease*, 10(2), 1133–1137. <https://doi.org/10.21037/jtd.2018.01.134>

- Bharuthram, S. (2012). Making a case for the teaching of reading across the curriculum in higher education. *South African Journal of Education*, 32(2), 205–214.
<https://doi.org/10.15700/saje.v32n2a557>
- Bharuthram, S. (2017). The reading habits and practices of undergraduate students at a higher education institution in South Africa: A case study. *The Independent Journal of Teaching and Learning*, 12(1), 50–62.
- Broekhuizen, H., & Spaull, N. (2017). *The 'Martha Effect': The compounding female advantage in South African higher education* [Working paper].
<https://resep.sun.ac.za/wp-content/uploads/2018/07/wp142017.pdf>
- Campisi, G., La Rocca, L., & Muzzioli, S. (2023). Assessing skewness in financial markets. *Statistica Neerlandica*, 77(1), 48–70. <https://doi.org/10.1111/stan.12273>
- Canova, S., Cortinovis, D. L., & Ambrogi, F. (2017). How to describe univariate data. *Journal of Thoracic Disease*, 9(6), 1741–1743.
<https://doi.org/10.21037/jtd.2017.05.80>
- Chalton, D. O., & Troskie, C. G. (1992). Q plots, a graphical aid for regression analysis. *Communications in Statistics - Theory and Methods*, 21(3), 625–636.
<https://doi.org/10.1080/03610929208830803>
- Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example* (5th ed.). Wiley.
- Cliff, A. (2014). Entry-level students' reading abilities and what these abilities might mean for academic readiness. *Language Matters*, 45(3), 313–324.
<https://doi.org/10.1080/10228195.2014.958519>
- Cliff, A. (2024). *Thirty years of literacies testing at the University of Cape Town: A critical reflection on the work and its impact*. Springer Nature.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.
<https://doi.org/10.2307/3587951>

- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 49(2), 222–251.
<https://doi.org/doi:10.3102/00346543049002222>
- Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In N. H. Hornberger (Ed.), *Encyclopedia of language and education* (2nd ed., Vol. 2, pp. 71–83). Springer US. https://doi.org/10.1007/978-0-387-30424-3_36
- Dagada, M. (2022). Foundation phase teachers' challenges in teaching reading in South Africa. *South African Journal of Education*, 42(Supplement 1), 1–9.
<https://doi.org/10.15700/saje.v42ns1a2219>
- Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of Learning Disabilities*, 39(6), 507–514.
<https://doi.org/10.1177/00222194060390060301>
- Department of Basic Education. (2023). *PIRLS 2021: South African preliminary highlights report*. Department of Basic Education.
- Department of Higher Education and Training. (2016). *2000 to 2008 first time entering undergraduate cohort studies for public higher education institutions*. Department of Higher Education and Training.
- Department of Higher Education and Training. (2021). *PrimTEd teaching reading study guides—Study guide 6: Vocabulary*. Department of Higher Education and Training.
- DiStefano, C., Shi, D., & Morgan, G. B. (2021). Collapsing categories is often more advantageous than modeling sparse data: Investigations in the CFA framework. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(2), 237–249.
<https://doi.org/10.1080/10705511.2020.1803073>
- Dong, Y., Tang, Y., Chow, B. W. Y., Wang, W., & Dong, W. Y. (2020). Contribution of vocabulary knowledge to reading comprehension among Chinese students: A meta-analysis. *Frontiers in Psychology*, 11, 525369.
<https://doi.org/10.3389/fpsyg.2020.525369>

- Dos Reis, K., Venter, A., & McGhie, V. (2019). Are high school teachers and university academics on the same page? A case of the university readiness of business education learners at two schools and two universities. *Journal of Education*, 76, 166–184. <https://doi.org/10.17159/2520-9868/i76a09>
- Dyers, C., & Antia, B. E. (2019). Multilingual and multimodal mediation in one university module: The people and processes involved. *Southern African Linguistics and Applied Language Studies*, 37(1), 62–76. <https://doi.org/10.2989/16073614.2019.1609365>
- Ferré, J. (2009). Regression diagnostics. In *comprehensive chemometrics* (pp. 33–89). <https://doi.org/10.1016/B978-044452701-1.00076-4>.
- Fouché, I. (2024). Measuring the impact of academic literacy interventions: Refining an evaluation design through self-reflection and feedback. *Applied Linguistics*, 44(4), 718–746. <https://doi.org/10.1093/applin/amad025>
- Fourie, C. M. (2020). Risk factors associated with first-year students' intention to drop out from a university in South Africa. *Journal of Further and Higher Education*, 44(2), 201–215. <https://doi.org/10.1080/0309877X.2018.1527023>
- Frankel, K. K., Becker, B. L. C., Rowe, M. W., & Pearson, P. D. (2016). From “What is reading?” to What is literacy? *Journal of Education*, 196(3), 7–17. <https://doi.org/10.1177/002205741619600303>
- Gholaminejad, R. (2020). A critical review of the word classification system. *HOW Journal*, 27(2), 156–167. <https://doi.org/10.19183/how.27.2.554>
- Goundar, S. (2012). Chapter 3: Research methodology and research method. In *Cloud Computing*. Research Gate Publications.
- Government Communication and Information System (GCIS). (2022). Official guide to South Africa 2021/22. *Government Communication and Information System (GCIS)*. <https://www.gcis.gov.za/sites/default/files/04%20People%20of%20South%20Africa%202021-22.pdf>
- Hadjerrouit, S., & Nnagbo, Celestine. I. (2021). Exploring numbas formative feedback for teaching and learning mathematics: an affordance theory perspective. *18th*

International Conference Cognition and Exploratory Learning in Digital Age 2021, 261–268. https://doi.org/10.33965/celda2021_202108L032

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson.

Hermida, Dr. J. (2009). The importance of teaching academic reading skills in first-year university courses. *SSRN Journal*, 1–12. <https://doi.org/10.2139/ssrn.1419247>

Hoang, G. T. H., Baten, K., De Cuypere, L., Hoang, T. T., & Taverniers, M. (2021). An exploratory study of predictors of vocabulary knowledge of Vietnamese preschool-age children in a city: Predictors of vocabulary knowledge in Vietnamese preschoolers. *Dutch Journal of Applied Linguistics*, 10, 1–24. <https://doi.org/10.51751/dujal9538>

Howie, S., Combrinck, C., Roux, K., Tshele, M., Mokoena, G., & Palane, N. M. (2017). *PIRLS Literacy 2016: Progress in international reading literacy study 2016: South African children's reading literacy achievement*. Centre for Evaluation and Assessment.

Huang, L., Ouyang, J., & Jiang, J. (2022). The relationship of word processing with L2 reading comprehension and working memory: Insights from eye-tracking. *Learning and Individual Differences*, 95, 1–12. <https://psycnet.apa.org/doi/10.1016/j.lindif.2022.102143>

Kavé, G. (2024). Vocabulary changes in adulthood: Main findings and methodological considerations. *International Journal of Language & Communication Disorders*, 59(1), 58–67. <https://doi.org/10.1111/1460-6984.12820>

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, 68(8), 1665–1692. <https://doi.org/10.1080/17470218.2015.1022560>

Khanyile, S., & Awung, F. (2023). Challenges of mother-tongue education in isiZulu: A case study of selected schools in the uThungulu district of Kwazulu-Natal. *African Journal of Inter/Multidisciplinary Studies*, 5(1), 1–8. <https://doi.org/10.51415/ajims.v5i1.1069>

- Kilic, M. (2019). Vocabulary knowledge as a predictor of performance in writing and speaking: A case of Turkish EFL learners. *PASAA*, 57(1), 133–164.
<https://doi.org/10.58837/CHULA.PASAA.57.1.6>
- Kim, Y. S. G. (2020). Toward integrative reading science: The direct and indirect effects model of reading. *Journal of Learning Disabilities*, 53(6), 469–491.
<https://doi.org/10.1177/0022219420908239>
- Kiszcza, A. (2019). Reading as a core component of developing academic literacy skills in L2 settings. *Lublin Studies in Modern Languages and Literature*, 43(3), 89–98.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, 54(4), 1076–1085.
<https://doi.org/10.1002/tesq.3004>
- Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, 21(2), 202–226.
<https://doi.org/10.1191/0265532204lt277oa>
- Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1), 33–51. <https://doi.org/10.1177/026553229901600103>
- Li, Y., Zhang, Y., Yu, M., & Li, X. (2019). Drawing and studying on histogram. *Cluster Computing*, 22(S2), 3999–4006. <https://doi.org/10.1007/s10586-018-2606-0>
- Loy, A., Follett, L., & Hofmann, H. (2015). Variations of Q-Q plots—The power of our eyes! *The American Statistician*, 70, 202–214.
<https://doi.org/10.1080/00031305.2015.1077728>
- Luft, J. A., Jeong, S., Idsardi, R., & Gardner, G. (2022). Literature reviews, theoretical frameworks, and conceptual frameworks: An introduction for new biology education researchers. *CBE—Life Sciences Education*, 21(3), 1–10.
<https://doi.org/10.1187/cbe.21-05-0134>
- Lundhede, T. H., Olsen, S. B., Jacobsen, J. B., & Thorsen, B. J. (2009). Handling respondent uncertainty in Choice Experiments: Evaluating recoding approaches against explicit modelling of uncertainty. *Journal of Choice Modelling*, 2(2), 118–147.
[https://doi.org/10.1016/S1755-5345\(13\)70007-1](https://doi.org/10.1016/S1755-5345(13)70007-1)

- Mahlangu, S. M., & Mtshali, T. I. (2024). Factors contributing to poor literacy among grade 4 learners in Gauteng province of South Africa. *Journal of China University of Mining and Technology*, 29(2), 77–82.
- Mangen, A., & Van Der Weel, A. (2016). The evolution of reading in the age of digitisation: An integrative framework for reading research. *Literacy*, 50(3), 116–124. <https://doi.org/10.1111/lit.12086>
- Margolin, S. J., Driscoll, C., Toland, M. J., & Kegler, J. L. (2013). E-readers, computer screens, or paper: Does reading comprehension change across media platforms? *Applied Cognitive Psychology*, 27(4), 512–519. <https://doi.org/10.1002/acp.2930>
- Mart, C. T. (2012). Developing speaking skills through reading. *International Journal of English Linguistics*, 2(6), 91–96. <https://doi.org/10.5539/ijel.v2n6p91>
- Matjila, D., & Pretorius, E. (2004). Bilingual and biliterate? An exploratory study of grade 8 reading skills in Setswana and English. *Per Linguam*, 20(1), 1–21. <https://doi.org/10.5785/20-1-77>
- Matope, S., & Senye-Awudi, P. (2024). Using the Strategy Instructions Approach (SIA) in teaching reading comprehension: A case study schools in the Amathole East District South in Africa. *E-Journal of Humanities, Arts and Social Sciences*, 4(13), 1581–1592. <https://doi.org/10.38159/ehass.202341316>
- McDonald, B. (2002). A teaching note on Cook's Distance – A guideline. *Research Letters in the Information and Mathematical Sciences*, 3, 127–128.
- Miles, J. (2014). Tolerance and variance inflation factor. *Wiley StatsRef: Statistics Reference Online*, 1–2.
- Millin, T. J. (2015). Reading to learn: A literature review within a South African context. *Stellenbosch Papers in Linguistics*, 44, 105–124. <https://doi.org/10.5774/44-0-170>
- Moll, I., Dlamini, R., Ndlovu, N., Drennan, G., Nkambule, F., & Phakathi, N. (2022). A developing realist model of the pedagogical affordances of ICTs. *South African Computer Journal*, 34(2), 50–75. <https://doi.org/10.18489/sacj.v34i2.1076>
- Moloi, K. (2019). Learners and educators as agents of social transformation in dysfunctional South African schools. *South African Journal of Education*, 39(1), 1–8. <https://doi.org/10.15700/saje.v39ns1a1800>
- Molotja, T. W. (2020). An Exploration of the academic reading strategies of first year English Education students at a South African university. *Journal of African Education*, 1(2), 103–123. <https://doi.org/10.31920/2633-2930/2020/1n2a5>

- Montero-SaizAja, A. (2021). Gender-based differences in EFL learners' language learning strategies and productive vocabulary. *Theory and Practice of Second Language Acquisition*, 7(2), 83–107. <https://doi.org/10.31261/TAPSLA.8594>
- Msimango, S., & Motala, S. (2023). Gendered 'leaky pipelines' in higher education: Towards a supervision framework for breaking through the glass ceiling. *Journal for New Generation Sciences*, 21(1), 116–131. <https://doi.org/10.47588/jngs.2023.21.01.a9>
- Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, 47(1), 91–108. <https://doi.org/10.1002/RRQ.011>
- Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, 5(1), 12–25.
- Nation, P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading*, 1, 5–16.
- National Benchmark Tests. (2023). *The National Benchmark Tests national report—2022 intake cycle*. <https://nbt.uct.ac.za/sites/default/files/2022%20INTAKE%20ANNUAL%20NATIONAL%20BENCHMARK%20TEST%20FINAL%2011072023.pdf>
- National Benchmark Tests. (2024). *Interpreting your results*. <https://www.nbt.ac.za/content/interpreting-your-results>
- Nel, C., Dreyer, C., & Kopper, M. (2004). An analysis of the reading profiles of first-year students at Potchefstroom University: A cross-sectional study and a case study. *South African Journal of Education*, 24(1), 95–103.
- Nel, C., & Nel, C. (2014). *A 3-tier model for supporting reading literacy among first year students*. 1–9. https://www.researchgate.net/figure/A-3-Tier-Reading-Literacy-Model-for-Supporting-First-Year-Students_fig1_265809914
- Neuman, S. B., & Dwyer, J. (2009). Missing in action: Vocabulary instruction in pre-K. *The Reading Teacher*, 62(5), 384–392. <https://doi.org/10.1598/RT.62.5.2>
- Nguyen, T., Khadka, R., Phan, N., Yazidi, A., Halvorsen, P., & Riegler, M. A. (2023). Combining datasets to increase the number of samples and improve model fitting. *2023 International Joint Conference on Neural Networks*. <https://doi.org/10.48550/arXiv.2210.05165>
- Ngwenya, T. (2010). Correlating first-year law students' profile with the language demands of their content subjects. *Per Linguam*, 26(1), 74–99. <https://doi.org/10.5785/26-1-15>
- Nizonkiza, D., & Van Dyk, T. (2015). Academic literacy of South African higher education level students: Does vocabulary size matter? *Stellenbosch Papers in Linguistics*, 44, 147–174. <https://doi.org/10.5774/44-0-159>

- Nolan, J. P. (2012). R as a tool in computational finance. In J.-C. Duan, W. K. Härdle, & J. E. Gentle (Eds.), *Handbook of computational finance* (pp. 781–804). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-17254-0_29
- Nortje, M. J. (2017). The effect of poverty on education in South Africa. *Educator Multidisciplinary Journal*, 1(1), 47–62.
- Nowell, L., Paolucci, A., Dhingra, S., Jacobsen, M., Lorenzetti, D. L., Lorenzetti, L., & Oddone-Paolucci, E. (2022). Interdisciplinary mixed methods systematic reviews: Reflections on methodological best practices, theoretical considerations, and practical implications across disciplines. *Social Sciences & Humanities Open*, 6(1), 1–8. <https://doi.org/10.1016/j.ssaho.2022.100295>
- Olawale, S. R., Chinagozi, O. G., & Ononokpono, N. J. (2023). Exploratory research design in management science: A review of literature on conduct and application. *International Journal of Research and Innovation in Social Science*, VII(IV), 1384–1395. <https://doi.org/10.47772/IJRISS.2023.7515>
- Owen-Smith, M. (2010). The language challenge in the classroom: A serious shift in thinking and action is needed. *Focus*, 56, 31–37.
- Parliament of the Republic of South Africa. (2023, May 2). *The NA approves South African Sign Language as the 12th official language*. <https://www.parliament.gov.za/press-releases/na-approves-south-african-sign-language-12th-official-language>
- Pellizzari, M., & Billari, F. C. (2012). The younger, the better? Age-related differences in academic performance at university. *Journal of Population Economics*, 25(2), 697–739. <https://doi.org/10.1007/s00148-011-0379-3>
- Pennsylvania State University. (2018). *Lesson 9: Influential points*. <https://online.stat.psu.edu/stat462/node/87/>
- Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *The Modern Language Journal*, 102(4), 713–731. <https://doi.org/10.1111/modl.12509>
- Posel, D., Hunter, M., & Rudwick, S. (2020). Revisiting the prevalence of English: Language use outside the home in South Africa. *Journal of Multilingual and Multicultural Development*, 43(8), 1–13. <https://doi.org/10.1080/01434632.2020.1778707>
- Posel, D., & Zeller, J. (2016). Language shift or increased bilingualism in South Africa: Evidence from census data. *Journal of Multilingual and Multicultural Development*, 37(4), 357–370. <https://doi.org/10.1080/01434632.2015.1072206>
- Pretorius, E. J. (2000). Reading and the UNISA student: Is academic performance related to reading ability? *Progressio*, 22(2), 35–48.

- Pretorius, E. J. (2002). Reading ability and academic performance in South Africa: Are we fiddling while Rome is burning? *Language Matters*, 33(1), 169–196. <https://doi.org/10.1080/10228190208566183>
- Pretorius, E. J. (2005). What do students do when they read to learn? Lessons from five case studies. *South African Journal of Higher Education*, 19(4), 790–812. <https://doi.org/10.4314/sajhe.v19i4.25668>
- Pretorius, E. J., & Klapwijk, N. M. (2016). Reading comprehension in South African schools: Are teachers getting it, and getting it right? *Per Linguam*, 32(1), 1–20. <https://doi.org/10.5785/32-1-627>
- Rabiner, D. L., Godwin, J., & Dodge, K. A. (2016). Predicting academic achievement and attainment: The contribution of early academic skills, attention difficulties, and social competence. *School Psychology Review*, 45(2), 250–267. <https://doi.org/10.17105/SPR45-2.250-267>
- Ranganathan, P., & Gogtay, N., J. (2019). An introduction to statistics – Data types, distributions and summarizing data. *Indian Journal of Critical Care Medicine*, 23(S2), 169–170. <https://doi.org/10.5005/jp-journals-10071-23198>
- Robertson, S. A., & Graven, M. (2020). A mathematics teacher's response to a dilemma: 'I'm supposed to teach them in English but they don't understand'. *South African Journal of Childhood Education*, 10(1), 1–11. <https://doi.org/10.4102/sajce.v10i1.800>
- Romadlon, F. N. (2017). Promoting students' reading proficiency through reciprocal technique. *Vision: Journal for Language and Foreign Language Learning*, 6(1), 50–60. <https://doi.org/10.21580/vjv6i11584>
- Rousseaux, E., Bolano, D., & Ritschard, G. (2013). *The Rsocialdata package: Handling survey data in R*. 1–23. <https://doi.org/10.13140/RG.2.1.4752.5527>
- Rubin, M., Scevak, J., Southgate, E., Macqueen, S., Williams, P., & Douglas, H. (2018). Older women, deeper learning, and greater satisfaction at university: Age and gender predict university students' learning approach and degree satisfaction. *Journal of Diversity in Higher Education*, 11(1), 82–96. <https://doi.org/10.1037/dhe0000042>
- Scheepers, R. (2016). The importance of vocabulary at tertiary level. *Journal for Language Teaching*, 50(1), 53–77. <https://doi.org/10.4314/jlt.v50i1.3>
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.
- Sebolai, K. (2022). The academic development approach to academic literacy in higher education South Africa: A disconnect between teaching and assessment. *Journal for Language Teaching*, 56(2), 1–16. <https://doi.org/10.56285/jltVol56iss2a5387>

- Sengupta, S. (2002). Developing academic reading at tertiary level: A longitudinal study tracing conceptual change. *The Reading Matrix*, 2(1), 1–37.
- Severance, C. (2024). *Python for everybody*. LibreTexts.
- Sibanda, J. (2017). Language at the grade three and four interface: The theory-policy-practice nexus. *South African Journal of Education*, 37(2), 1–9. <https://doi.org/10.15700/saje.v37n2a1287>
- Singh, A. (2021). An introduction to experimental and exploratory research. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3789360>
- South African Government News Agency. (2023, October 12). *92.1% of SA population owns a cellphone*. South African Government News Agency. <https://www.sanews.gov.za/south-africa/921-sa-population-owns-cellphone>
- Spjeldnæs, K., & Karlsen, F. (2024). How digital devices transform literary reading: The impact of e-books, audiobooks and online life on reading habits. *New Media & Society*, 26(8), 4808–4824. <https://doi.org/10.1177/14614448221126168>
- Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139–152. <https://doi.org/10.1080/09571730802389975>
- Stats SA. (2024, March 26). *Marginalised communities in SA achieve new heights in education*. Statistics South Africa. <https://www.statssa.gov.za/?p=17134#:~:text=Specifically%2C%2013%2C1%25%20of,and%20promote%20inclusive%20educational%20policies.&text=advanced%20degrees%2C%20with%20males%20consistently,download%20the%20full%20report%20here>
- Stein, N. (2017). Chapter 11: Language in schools. In *Basic education rights handbook - Education rights in South Africa*. SECTION27.
- Stoffelsma, L., & Spooren, W. (2019). The relationship between English reading proficiency and academic achievement of first-year science and mathematics students in a multilingual context. *International Journal of Science and Mathematics Education*, 17(5), 905–922. <https://doi.org/10.1007/s10763-018-9905-z>
- Sutton, C. E. (2019). *Does age matter in higher education? Investigating mature undergraduate students' experiences* [PhD Thesis, University of Leicester]. <https://doi.org/10.25392/leicester.data.10321283.v1>
- Swaraj, A. (2019). Exploratory research: Purpose and process. *Parisheelan Journal*, 15(2), 665–670.

- Taylor, S., & Von Fintel, M. (2016). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach. *Economics of Education Review*, 50, 75–89. <https://doi.org/10.1016/J.ECONEDUREV.2016.01.003>.
- Temko, E. (2024). *The stories behind social statistics: Data analysis, interpretation, and communication*. LibreTexts.
- Terasawa, T. (2024). Relationship between English proficiency and socioeconomic status in Asia: Quantitative cross-national analysis. *World Englishes*, 1–44. <https://doi.org/10.1111/weng.12705>
- UNESCO Institute for Statistics. (2025). *Glossary: Home language*. <https://uis.unesco.org/en/glossary-term/home-language>
- United Nations Educational, Scientific and Cultural Organisation. (2023, November 8). *Gender equality in and through the teaching profession*. UNESCO. <https://www.unesco.org/en/articles/gender-equality-and-through-teaching-profession#:~:text=Teaching%20is%20increasingly%20a%20female%2Ddominated%20profession%E2%80%A6&text=But%20at%20higher%20levels%20of%20education%2C%20their%20numbers%20dwindle>.
- Van Der Merwe, D. (2018). Aspects of academic language proficiency of intermediate phase teacher education students. *South African Journal of Childhood Education*, 8(1), 1–9. <https://doi.org/10.4102/sajce.v8i1.555>
- Van Dyk, T., & Van De Poel, K. (2013). Towards a responsible agenda for academic literacy development: Considerations that will benefit students and society. *Journal for Language Teaching*, 47(2), 43–69. <https://doi.org/10.4314/jlt.v47i2.3>
- Van Schalkwyk, S. (2008). *Acquiring academic literacy: A case of first-year extended degree programme students* [PhD Thesis, Stellenbosch University]. <https://www.tandfonline.com/doi/full/10.2989/SALALS.2009.27.2.6.869>
- Van Staden, S., Bosker, R., & Bergbauer, A. (2016). Differences in achievement between home language and language of learning in South Africa: Evidence from prePIRLS 2011. *South African Journal of Childhood Education*, 6(1), 1–10. <https://doi.org/10.4102/sajce.v6i1.441>
- Van Voorhis, C. R. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, 3(2), 43–50. <https://doi.org/10.20982/tqmp.03.2.p043>
- Wampold, B. E., & Freund, R. D. (1987). Use of multiple regression in counseling psychology research: A flexible data-analytic strategy. *Journal of Counseling Psychology*, 34(4), 372–382. <https://doi.org/10.1037/0022-0167.34.4.372>

- Webb, S. A., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal*, 44(3), 263–277. <https://doi.org/10.1177/0033688213500582>
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated vocabulary levels test: Developing and validating two new forms of the VLT. *International Journal of Applied Linguistics*, 168(1), 33–69. <https://doi.org/10.1075/itl.168.1.02web>
- Weder, N. (2022). *The effect of different interventions on the inferencing abilities of academically vulnerable first-year students* [PhD Thesis, University of Pretoria]. <https://rgdoi.net/10.13140/RG.2.2.34801.04961>
- Weine, E., McPeck, M. S., & Abney, M. (2023). Application of equal local levels to improve Q-Q plot testing bands with R package qqconf. *Journal of Statistical Software*, 106(10), 1–31. <https://doi.org/10.18637/jss.v106.i10>
- Wills, G., Ardington, C., Pretorius, E., & Sebaeng, L. (2022). *Benchmarking early grade reading skills: English First Additional Language. Summary report.* (pp. 1–17). Khulisa Management Services.
- Wong, A. S. C., & Lee, J. Y. V. (2020). Investigating the predictive role of vocabulary in written performance. *Borneo Akademika*, 4(4), 70–83. <https://doi.org/10.24191/BAv4i4/80759>

Appendix A: Ethical clearance



HUMAN RESEARCH ETHICS COMMITTEE (NON-MEDICAL)

Registration number: REC-101114-044

23 October 2024

SCHOOL OF SOCIAL SCIENCES ETHICS COMMITTEE

Ethics clearance/waiver number: WSOSSINTR092024/7

Re: Memory Ponela (1489805)

To whom it may concern:

Memory Ponela (1489805) is currently registered as a Master's student in the School of Social Sciences, University of the Witwatersrand, Johannesburg.

This letter is to confirm that, at the time of writing, Memory does not need ethical clearance for the study entitled "Exploring the Predictors of Academic Vocabulary Proficiency among First-Year University Students".

This decision has been reached based upon a description of the project supplied by Memory to the School of Social Sciences Ethics Committee, constituted as a subcommittee of the University Human Research Ethics Committee (Non-Medical), which has been evaluated by the subcommittee chair. This decision has then been ratified by the University Human Research Ethics Committee (Non-Medical).

If, however, Memory changes the methods of data collection and analysis for this project, this decision may no longer be valid. If such changes take place, this should be communicated to the School of Social Sciences Ethics Committee.

Please feel free to contact me should you require any further information.

Thank you.

Yours sincerely,

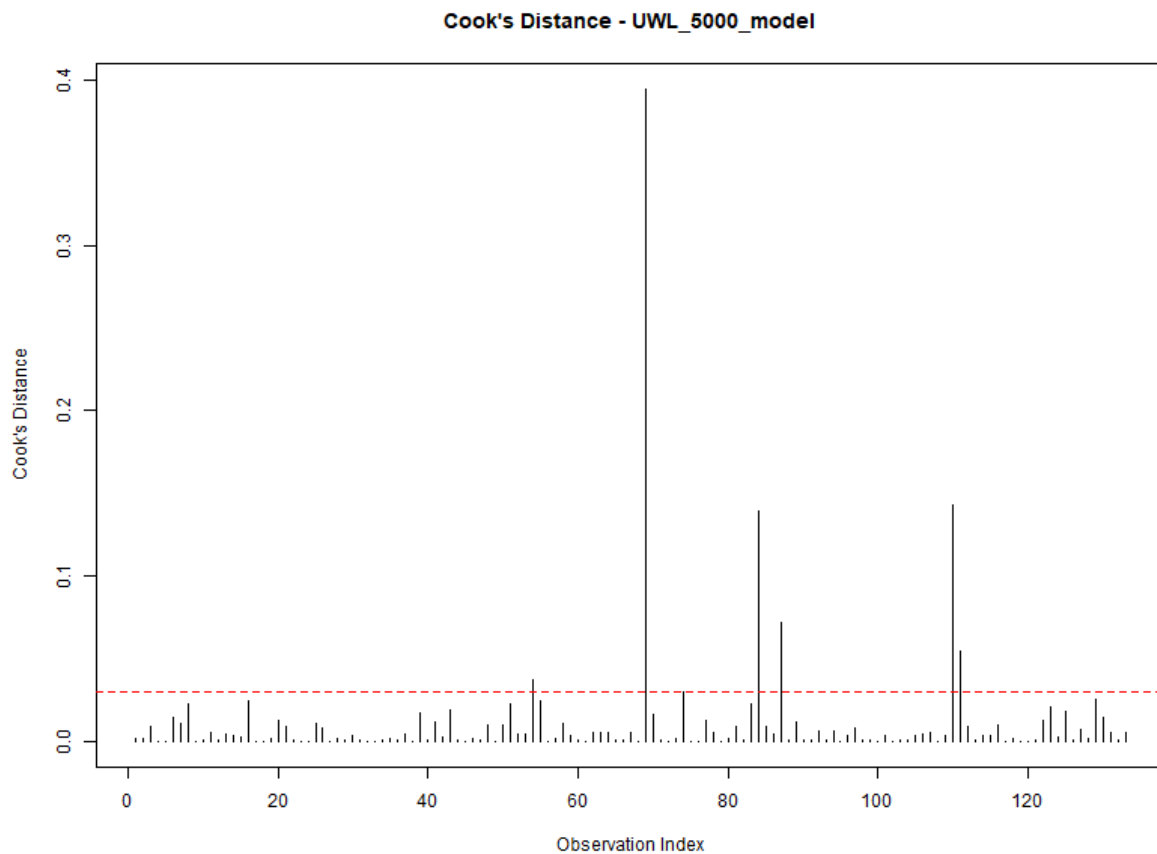
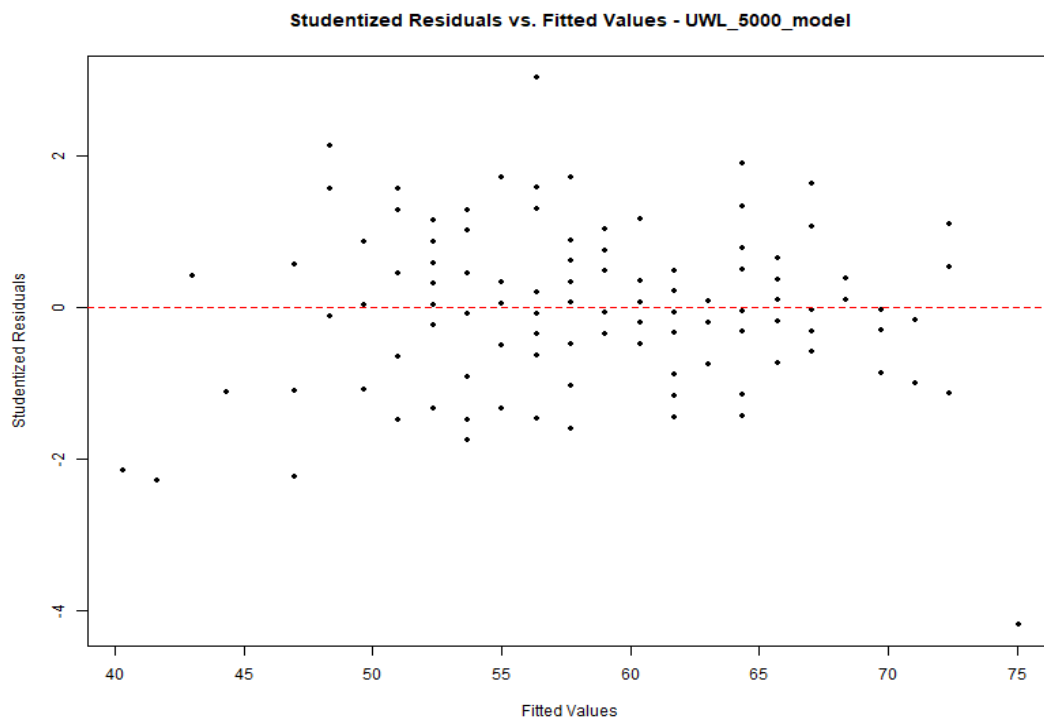


Dr. Celso Monjane
Co-chair, School of Social Sciences Ethics Committee

cc: Dr. Ilse Fouche and Dr. Laura Drennan (supervisor)

Shaun Schoeman (Senior Administrative Officer)
Solomon Mahlangu House, 10th Floor, Room 10004, Jorissen Street, Braamfontein, Johannesburg
Private Bag 3, Wits 2050
T + 27(0)11 717 1408 | E Shaun.Schoeman@wits.ac.za | hrec-medical.researchoffice@wits.ac.za
www.wits.ac.za/research/about-our-research/ethics-and-research-integrity/

Appendix B: Diagnostics results - UWL and 5000-word level simple regression



Appendix C: Diagnostics results - Multiple regression

