New Measures Of Academic Collocation Knowledge: Wordlist-Based Test Development And Argument-Based Validation

BY

NGUYEN THI MY HANG

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the requirements for the degree of
Doctor of Philosophy

Victoria University of Wellington
2022

Dedication

To my amazing parents, Nguyen Tien Dung and Dang Thi Thanh Thuy, for their unconditional love and support, not only for me but also for my two little kids.

To my husband – Nguyen Vo Dao, my daughter – Nguyen Bach Duong (Kiwi) and my son – Nguyen Nhat Nam (Apec), for accepting the challenge of being separated in three different countries. While my husband and I were doing our PhDs in Japan and New Zealand, the kids were looked after by my parents in Vietnam.

To my family-in-law for always being willing to assist us when we needed help.

Acknowledgements

First and foremost, I am profoundly grateful to my primary supervisor, Professor Averil Coxhead, for her constant support in both academic and personal matters. With her trust, encouragement and guidance, my PhD roller coaster ride was guaranteed. I am grateful for the knowledge and skills she has always been so eager to share, as well as her wonderful sense of humour and positive attitude. Working with her is the key factor that makes my PhD journey fulfilled with motivation and joy.

I would like to express my sincere appreciation to my secondary supervisor, Associate Professor Peter Gu, for sharing his vast knowledge of language testing and for his willingness to clearly explain any concept in this field to me, no matter how abstract or technical it may be. I am also thankful for his availability to sit and talk whenever I needed help, despite his busy schedule.

I wish to thank all the people whose assistance was a milestone in the completion of this research. These include the authors of the two academic collocation lists: Kirsten Ackermann, Yu-Hua Chen, Lei Lei and Dilin Liu for sharing their lists to be used in this thesis, my colleagues in Vietnam for their tremendous support for my data collocation during the challenging time of COVID-19, my participants in Vietnam and New Zealand for their time and enthusiasm, and the learning advisors at Victoria University of Wellington for their useful advice on my writing and statistics skills.

I gratefully acknowledge the financial support from Victoria University of Wellington in the form of a Victoria Doctoral Scholarship for tuition and living allowance, as well as a Faculty Research Grant for data collection.

It should be acknowledged that parts of Section 3.1.1 (Chapter 3), Sections 4.1 and 4.2 (Chapter 4) have been adapted for a joint authored article (with Professor Averil Coxhead as the second author) titled “Evaluating multiword units word lists for academic purposes”, which has been accepted for publication by the *ITL International Journal of Applied Linguistics*. I am grateful to the editor and anonymous reviewers of this journal for their useful feedback on the article.

I would also like to express my gratitude to my three examiners: Associate Professor Irina Elgort (Victoria University of Wellington), Emeritus Professor John Read (The University of Auckland) and Dr. Ron Martinez (University of California) for their insightful comments and suggestions.

Last but not least, I would like to give my warmest thanks to all the people in Vietnam and New Zealand who encouraged and supported me as I worked on this thesis, including family members, friends, staff at Victoria University of Wellington and colleagues at the University of Foreign Language Studies in Danang, Vietnam.

Abstract

Knowledge of academic vocabulary is essential for second or foreign language learners who are preparing for studying at English-medium universities. This vocabulary comprises single words (e.g., *approximate, component* and *establish*) and multiword units, including frequent two-word academic collocations. These items occur across different disciplines (e.g., *ultimate goal* and *key element*) and have been identified through corpus-based research which has resulted in several word lists (e.g., Ackermann & Chen, 2013; Lei & Liu, 2018). While there are tests which target knowledge of general academic vocabulary based on single word lists, there is a lack of tests of academic collocation knowledge. Such assessment is beneficial for troubleshooting problems with academic collocations at an early stage so that support for learning these items can be provided in a timely manner. This study aims to fill this gap by developing and validating two separate measures of recognition and recall knowledge of general academic collocations for diagnostic purposes.

To that end, this research first adapted an existing framework from Nation (2016) to evaluate two published lists of academic collocations. The evaluation was to select the most representative items for developing two Academic Collocation Tests (ACTs): the recognition test (multiple-choice format) and the recall test (gap-fill format). The test development was guided by an evidence-centred design framework (Mislevy & Yin, 2013). The validation process then employed an argument-based approach (Kane, 2013) to collect validity evidence. A total of 343 tertiary students (233 in Vietnam and 110 in New Zealand) took part in this study. They completed a background questionnaire which included demographic information and language proficiency (e.g., IELTS scores and learning experience). They took both of the ACTs and also completed the Vocabulary Size Test (VST) (Nation & Beglar, 2007), which was used as a measure of general vocabulary knowledge. Forty-four of the participants took part in a post-test interview to share their reflections on the tests and re-took the ACTs verbally for the assessment of test-retest reliability. Data gathering took place via online platforms because of the COVID-19 pandemic.

Five main findings arose from this thesis. First, results of the wordlist evaluation process indicated that the Academic Collocation List (Ackermann & Chen, 2013) provided the best

source of items for testing purposes in the present study. Second, statistical analyses showed that test items developed from that list worked well together to measure the intended construct. Third, reflections from test-takers revealed that the ACTs allowed them to demonstrate their knowledge of academic collocations, although the online test-taking condition was not ideal. Fourth, the ACTs were found to be highly reliable, as evidenced by high reliability indices. Finally, scores on the ACTs were positively correlated with scores on other tests of similar constructs, including the VST and IELTS. The relationship between ACT scores and time spent studying English was also significant. That said, ACT scores were not significantly correlated with the frequency of academic collocations and time spent studying in an English-speaking context.

Based on these findings, this thesis offers pedagogical implications to support English for Academic Purposes (EAP) teaching and learning with improving academic collocation knowledge. This study advances the field of vocabulary assessment by applying test development and validation frameworks to create rigorous tests for EAP. It also provides a model for evaluating word lists of multiword units, which lays the foundation for a similar practice in wordlist studies and supports the further application of wordlist-based test development.

Table of Contents

List of tables

List of figures

List of appendices

List of abbreviations

Abbreviation	Full form
ACL	Academic Collocation List (Ackermann & Chen, 2013)
AC Recall Test	Academic Collocation Recall Test
AC Recognition Test	Academic Collocation Recognition Test
AECL	Academic English Collocation List (Lei & Liu, 2018)
ACT(s)	Academic Collocation Test(s)
AVL	Academic Vocabulary List (Gardner & Davies, 2014)
BASE	British Academic Spoken English (Thompson & Nesi, 2001)
BAWE	British Academic Written English (Nesi & Gardner, 2012)
BNC	British National Corpus (BNC Consortium, 2007)
COCA	Corpus of Contemporary American English (Davies, 2008-2017)
EAP	English for Academic Purposes
EFL	English as a Foreign Language
ESL	English as a Second Language
FRQ	Frequency
IELTS	International English Language Testing System
Gram AECL	Grammatical collocations in the Academic English Collocation List (Lei & Liu, 2018)
GSL	General Service List (West, 1953)
L1	First language
L2	Second language
Lex AECL	Lexical collocations in the Academic English Collocation List (Lei & Liu, 2018)
MI	Mutual Information
MNSQ	Mean-square
PCAR	Principal component analysis of residuals
PICAE	Pearson International Corpus of Academic English (Ackermann et al., 2011)
RQ	Research question
SD	Standard Deviation
TLU	Target Language Use
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication
VLT	Vocabulary Levels Test (Schmitt et al., 2001)
VST	Vocabulary Size Test (Nation & Beglar, 2007)
ZSTD	Standardised Z

Chapter 1 Introduction

1.1 Rationale for the study

Knowledge of academic vocabulary in English is important for speakers of other languages who plan to undertake university-level studies in English (Coxhead, 2000; Nation, 2013). Most research into academic vocabulary has focused on single words (Coxhead, 2018), and through the development of general academic word lists (e.g., the Academic Word List by Coxhead, 2000; the Academic Vocabulary List by Gardner & Davies, 2014) and tests based on those lists (e.g., the Academic Vocabulary Test by Pecorari et al., 2019). More recently, attention in the field of vocabulary studies has focused on multiword units, such as collocations and idioms, whereby words are combined to create new meanings (Hinkel, 2018; Siyanova-Chanturia & Pellicer-Sánchez, 2019). The focus of this study is general academic English collocations, which are two-word combinations that frequently appear in a wide range of academic disciplines, such as *significant difference* and *key element*. Two-word collocations are the most frequent multiword units in academic texts, and they are typically the building blocks of larger multiword units (Biber & Barbieri, 2007).

This section first outlines the importance of testing knowledge of academic collocations, and then identifies the gaps in this assessment practice. These two key areas motivate the present study.

1.1.1 Importance of assessing academic collocation knowledge

For language recognition and production in academic settings, knowledge of academic collocations is crucial. Research has shown that there are a large number of academic collocations in written academic texts (Biber, 2006; Lei & Liu, 2018). Without knowledge of these words, English as a Foreign Language (EFL) and English as a Second Language (ESL) learners may struggle with their studies at English-medium universities. On the one hand, a lack of receptive knowledge of academic collocations can lead to miscomprehension when reading academic materials, especially when learners tend to break a collocation down into individual words and infer the meaning from its components (Wray, 2002; Nation, 2013). On the other hand, a lack of productive knowledge of academic collocations can lead to learners using

incorrect word combinations, or deviant collocations, which affects intelligibility and formality. Coxhead (2008) found that learners in an ESL context tend to focus on meaning and use of single academic words in their writing rather than collocations. Consequently, they tend to use unconventional word combinations in their writing because they focus on whether the phrases make sense semantically rather than considering whether the words might be collocates. For example, instead of using the collocation *central issue*, learners may use *centre issue*, which is not appropriate in academic writing.

There are several reasons why a measure of academic collocation knowledge is necessary. First, this knowledge is not measured by general language proficiency tests. English is widely used as a medium of instruction at the tertiary level in both English-speaking countries as well as non-native contexts (Hyland & Shaw, 2016). Assessing learners' English competence prior to admission to such educational settings has become common practice. Many universities use scores of international language proficiency tests such as the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) as pre-admission requirements. However, these tests do not assess vocabulary as a separate component. Second, an academic collocation test will be beneficial for teachers to troubleshoot learners' problems with academic collocations so that additional support could be provided. For example, teachers may use the results to decide whether to pay more attention to the development of academic collocation knowledge as part of an English for Academic Purposes (EAP) course. Third, there is still a lack of a rigorous measure of academic collocations. To the best of my knowledge, there have been only three studies that claimed to assess learners' knowledge of academic collocations (i.e., Frankenberg-Garcia, 2018; Voss, 2012; Wongkhan & Thienthong, 2020). Even though the previous tests make a valuable contribution to understanding the construct of collocation knowledge, none of them has been recognised as a standard. The following section will look at the issues of prior tests more closely.

1.1.2 Issues in assessing academic collocation knowledge

This section will briefly outline five main issues of collocation tests in earlier studies. These tests will later be discussed in detail in Chapter 2 (Section 2.4). Although general collocations are not

the focus of this study, their measures are included in the review to lay the groundwork for the current study.

First, some previous tests, such as those of Voss (2012) and Wongkhan and Thienthong (2020), regard collocation knowledge as a property of individual word knowledge rather than as an independent construct. That is, they measure only one word of a collocation pair instead of the entire collocation. For example, Voss (2012) focuses on verb + noun collocations (e.g., *collect data*) and requires test-takers to fill a gapped sentence with only one verb (e.g., *collect*).

Second, item representativeness is often not well justified in prior tests, which might be the result of a lack of systematic sampling from a credible source such as corpus-based word lists (Nation, 2016). A well-constructed word list is a representative sample of a corpus that captures actual language use. For example, Ackermann and Chen's Academic Collocation List (2013) contains written academic collocations which were selected from an academic corpus made up of articles, book chapters and textbooks. Word lists have been used as the foundation for developing tests of single academic words, such as the Academic Vocabulary Test (Pecorari et al., 2019) based on the Academic Vocabulary List (Gardner & Davies, 2014). They have also been used to design tests of multiword units, such as the Phrase Test developed by Martinez (2011) from his Phrasal Expressions List. However, this direction has not been explored with academic collocation tests.

The third issue is the absence of a test development framework in previous studies. This issue is closely bound to the first two issues. Because the steps in creating the tests are not guided by a comprehensive framework, the relationship between the target knowledge and the test tasks appears to have been overlooked during the test design process. For instance, the test of Wongkhan and Thienthong (2020) was designed with only ten items to measure learners' knowledge of academic collocations. No justifications were given on how the collocations were selected, how to ensure their representativeness, or how test-takers' performance on the test could demonstrate their knowledge of academic collocations. All of these considerations are taken into account within an evidence-centred design framework (Mislevy & Yin, 2013) for test development which has been employed for various language tests such as Test of English for International Communication (TOEIC) (e.g., Hines, 2010) or TOEFL (e.g., Pearlman, 2008), but not yet for vocabulary measures.

Fourth, earlier tests tend to place little emphasis on validity assessment. One risk is that validation evidence (if any) could include statistical analyses which may seem disjointed. Gyllstad (2009), for example, provided coefficient reliability (Cronbach's alpha > .89) and correlation with another vocabulary measure (r > .80) as evidence of validity for his COLLEX and COLLMATCH tests (see more in Section 2.4.1.1). However, Gyllstad (2009) did not explicitly state which validation framework was being used, which means readers are left to interpret what those results really mean in relation to the test validity. Within a coherent framework such as the argument-based validation (Kane, 2004, 2013), all research findings are able to be connected in a comprehensive network to assess the validity. This framework has been successfully applied to various measures of language proficiency such as TOEFL (e.g., Chapelle et al., 2008) and IELTS (e.g., Alavi et al., 2018; Aryadoust, 2011), but it has yet to be widely employed in the field of vocabulary assessment.

Last, there is a lack of measures to test both recognition and recall knowledge of the same target academic collocations so that these two types of knowledge can be directly compared. When looking into vocabulary knowledge, it is essential to make the distinction between receptive/ recognition and productive/ recall knowledge. The former refers to the ability to recognise the form-meaning link of a word, whereas the latter involves the ability to produce or retrieve a word to be used in a particular context (Nation, 2013; Read, 2000; Schmitt, 2010). The present study uses recognition/recall instead of receptive/productive terminologies because according to Schmitt and Schmitt (2020), "receptive/productive knowledge of vocabulary is usage-based, and should presumably be measured with skill-based instruments" (p.37). That is, lexical knowledge is embodied in the ability to comprehend key messages when listening or reading (receptive) and the ability to produce words appropriately when speaking or writing (productive). With recognition/recall knowledge, words can be tested in isolation using formats such as multiple-choice or fill-in-the-blank. The selection of the terms is important as they reflect the nature of the tests. Because recognition and recall refer to different aspects of word knowledge, separate measures are required to have a better insight into knowledge of academic collocations.

1.2 Aims and scope of the study

To address the aforementioned gaps in the assessment of academic collocations, the present study has two primary aims. The first aim is to develop two Academic Collocation Tests (ACTs) based on Ackermann and Chen's (2013) Academic Collocation List. These tests are suitable for EAP students who are studying or plan to study at an English-medium university. They can take the tests to diagnose their recognition and recall knowledge of two-word collocations in academic contexts. The results of the ACTs provide information for EAP teachers to determine whether students need more support in developing their knowledge of academic collocations. The ACTs are also potentially useful for research purposes or theory-building. For example, the tests can be used as pre and post tests in intervention studies. The steps of developing the ACTs are embedded in three layers of the evidence-centred design framework (Mislevy & Yin, 2013), including domain analysis, domain modelling and assessment implementation. The domain analysis involves a wordlist evaluation process to determine which of two academic collocation lists, Ackermann and Chen (2013) and Lei and Liu (2018), provides a better source of items for test development. The domain modelling includes the selection of academic collocation items and the creation of test items. The assessment implementation concerns piloting and finalising the tests.

The second aim of this study is to explore the interpretations of the ACT scores following the argument-based approach to validation (Kane, 1992; 2006; 2013). This involves gathering qualitative and quantitative data in two educational contexts (Vietnam and New Zealand) for the assessment of three main inferences made about the ACTs:

- Evaluation inference: Test-takers' performance on the ACTs is appropriately observed and scored.
- Generalisation inference: Test-takers' scores on the ACTs are consistent across tasks and testing occasions.
- Extrapolation inference: Performance on the ACTs is indicative of the target construct of academic collocations.

The scope of this study is limited to written academic collocations which occur across various disciplines. As there have already been several studies on general collocations (e.g., Barfield & Gyllstad, 2009; Nguyen & Webb, 2017), the present study aims to focus on a narrower group of frequent collocations in the academic domain. They do not include domain-specific collocations which are commonly used in only a particular area (e.g., *absolute altitude* and *hard landing* in aviation). The reason for this is that general academic collocations are beneficial for a wider group of learners compared to domain-specific collocations. Furthermore, this study is not concerned with spoken collocations. The correct use of collocations is especially important in academic writing because it is one of the most common forms of assessment at university and requires a high level of language accuracy. For those reasons, general collocations, domain-specific collocations and spoken academic collocations are beyond the scope of the present study.

1.3 Significance of the study

This study is valuable for a number of reasons. Firstly, it contributes to the field of vocabulary studies by providing two ready-to-use measures of academic collocations, the AC Recognition Test and the AC Recall Test, for pedagogical and research purposes. EAP teachers can use these tests to diagnose learners' current knowledge of academic collocations and set the learning goals. The tests are also useful research tools to investigate how knowledge of academic collocations develops over time.

Secondly, this study contributes to the field of vocabulary testing by employing the evidence-centred design framework (Mislevey & Yin, 2013) for test development and the argument-based framework for validation (Kane, 2004, 2013). These frameworks have been widely used in the language testing field to strengthen test design and provide a comprehensive evaluation of test quality. The current research can act as an example to illustrate how these frameworks work for tests of vocabulary, which generates motivation for test developers to apply them for other vocabulary measures.

Thirdly, the present study also contributes to the field of wordlist research as part of vocabulary studies through the evaluation and comparison between two existing lists of academic

collocations to select the best candidate for test development. This study provides a model for evaluating word lists of multiword units, which lays the foundation for a similar practice in wordlist studies and supports the further application of wordlist-based test development.

1.4 Organisation of the thesis

This thesis is structured as follows. Chapter 2 first introduces important concepts in the present study, including the definition of academic collocations and frameworks for test development and validation. It also provides an in-depth discussion and critique of research on the assessment of collocation knowledge, which leads to the research questions of the present study.

Chapter 3 describes the methods employed for the development and validation of the Academic Collocation Tests (ACTs). This chapter includes a comprehensive description of participants, materials, data collection and analysis.

Chapter 4 outlines the steps in the creation of the ACTs. The test development is based on the Academic Collocation List (Ackermann & Chen, 2013) following the evidence-centred design framework (Mislevy & Yin, 2013). This list is selected as a result of a wordlist evaluation process. The chapter then goes into detail about item selection, test format, writing test items and piloting.

Chapters 5 and 6 present the study findings in relation to the research questions and the three inferences in the argument-based validation framework of the ACTs: Evaluation, Generalisation and Extrapolation. Chapter 5 reports both qualitative and quantitative results which serve as evidence for the assessment of the Evaluation inference. Chapter 6 then reports the findings of the test reliability to examine the Generalisation inference and the correlation analyses to assess the Extrapolation inference.

Chapter 7 connects the findings from Chapters 4-6 and discusses them in three emerging themes: (1) developing academic collocation knowledge, (2) using word lists in creating tests of academic collocations and (3) applying the development and validation frameworks for academic collocation tests.

Finally, Chapter 8 concludes the thesis by highlighting the contributions of the current research, implications for different stakeholders (e.g., EAP teachers and test developers), research limitations, and directions for future studies. A brief reflection on my PhD journey completes Chapter 8 and the thesis as a whole.

Chapter 2 Literature review

This chapter explores the theories, concepts and frameworks to form the foundation for the development and validation of the Academic Collocation Tests (ACTs). It begins with the definition and importance of academic collocations in Sections 2.1 and 2.2. The next section explores the existing word lists on academic collocations as potential sources for the test development. Section 2.4 then identifies gaps in testing academic collocation knowledge and paves the way for the creation of the ACTs. The chapter continues with the development and validation frameworks employed for the ACTs in Section 2.5. Section 2.6 introduces the research questions that guide the process of collecting evidence for the validation of the ACTs. Finally, Section 2.7 summarises this chapter.

2.1 What are academic collocations?

This section begins with a discussion of collocations in relation to other categories of formulaic language such as phrasal verbs and idioms. The section then identifies common approaches to defining collocations and provides definition of academic collocations in the present study.

2.1.1 Collocations and other categories of formulaic language

In addition to a lexical single unit such as *number, fact* and *view*, English vocabulary also comprises preconstructed multiword combinations, or formulaic language such as *a large number of, as a matter of fact* and *point of view*. Wray (2002) defines the term formulaic language (also known as multiword expression or prefabricated language) as:

> "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (p.9).

Two important features of formulaic language related to the above definition are holistic processing and predictability. These two features are interrelated, with one leading to the other. Under the holistic view, multiword expressions such as *so far so good* or *break the ice* are stored

in the mental lexicon as a whole unit rather than as individual constituent words. This leads to the processing advantage in which formulaic language is supposed to be processed more quickly than random combinations (Carrol & Conklin, 2017, 2020; Gyllstad & Wolter, 2016). Meanwhile, the predictability refers to the ability of recalling the remaining part of a multiword expression based on the beginning of it (Siyanova-Chanturia & Martinez, 2015). For example, when hearing "sooner or …", one would easily fill in "later".

Formulaic language is an umbrella term to cover other linguistic phenomena, including collocations, lexical bundles, phrasal verbs and idioms. These four categories of formulaic language have received substantial attention in previous research (see Siyanova-Chanturia & Pellicer-Sánchez, 2019; Vilkaitė-Lozdienė, 2016; Wray, 2002). Understanding the differences between these categories helps to appropriately define the construct of academic collocations in the present study.

While collocations are two words that frequently occur together (see more in Section 2.1.2), lexical bundles are longer (three or more words) and completely fixed sequences, such as *there is a* or *as a result*. Biber et al. (1999) defined lexical bundles as "extended collocations: bundles of words that show statistical tendency to co-occur" (p.989). This means that lexical bundles are any group of words occurring together with high frequency, without further criteria of well-formedness or meaning completeness. Idioms, which are also made up of two or more words, have a non-compositional meaning (Vilkaitė-Lozdienė, 2016). In other words, the meaning of an idiom cannot be inferred from its constituent words (e.g., *once in a blue moon* means not very often). Phrasal verbs, on the other hand, are verb-preposition combinations which have an idiomatic meaning (Biber et al., 1999), for instance, *put up with* (i.e., tolerate or endure) or *get over* (i.e., recover from something). It should be noted that such classification of formulaic language categories is only relative. Sometimes a sequence can belong to more than one category (e.g., *on the other hand* can be listed as either a lexical bundle or an idiom). The focus on collocations in the present study is because they are more frequent than idiomatic sequences (i.e., phrasal verbs and idioms) and they are the building blocks of larger multiword units (e.g., lexical bundles) (Biber & Barbieri, 2007). The section that follows will look more closely into the definition of collocations.

2.1.2 Approaches to defining collocations

The two traditional approaches to identifying collocations are phraseology and statistics. In both approaches, collocations are two words that appear together more frequently than expected by chance (Biber et al., 1999). The phraseological approach (see Cowie, 1998; Howarth, 1996; Nesselhauf, 2005), however, adds a semantic layer to characterise collocations as restrictively co-occurring words and relative transparency in meaning (Laufer & Waldman, 2011). This implies that collocation components cannot be replaced for a synonym without loss of idiomatic meaning. For example, one would say *big surprise* but not *large surprise.* The statistics approach (e.g., Eyckmans, 2009; Nguyen & Webb, 2017), on the other hand, relies on objective measures such as frequency of co-occurrence and strength of association (e.g., mutual information score) to identify collocations. This method helps to point out collocations that are both frequent and strongly associated, and thus are likely to be useful for learners.

The advent of computers and the development of corpus linguistics has a transformative impact on how collocations are defined and identified. The two most prevalent approaches in corpus linguistics are quantitative and mixed methods approaches, both of which employ computer extraction to identify commonly paired words from a corpus (i.e., a collection of texts) based on specified statistical criteria. The quantitative approach (e.g., Durrant, 2009; Lei & Liu, 2018), which follows the traditional statistical approach, defines collocations as frequent co-occurrences, without any additional semantic criteria. The mixed methods approach, nevertheless, adds a layer of qualitative judgements to select collocations that meet certain purposes. In the study of Ackermann and Chen (2013), for example, an expert panel was employed to select pedagogically relevant collocations for EAP teachers and learners.

The present study follows the corpus-based mixed-methods approach to identifying academic collocations for testing purposes. Compared to the traditional approaches, the corpus-based identification of collocations allows researchers to focus on items that are more likely to be encountered or occur in a specific context (Rogers et al., 2021). As corpus-based research applies statistical criteria to identify collocations, the transparency of collocation meanings is often not considered. Subjective judgements are involved in the process of selecting collocations to match an intended purpose (e.g., pedagogical purpose) instead of judging the semantic transparency or

restrictiveness of the collocations. The following section discusses in depth the construct of academic collocations in the current research.

2.1.3 Defining the construct of academic collocations

Academic collocations in this study are defined as pairs of words that appear commonly together in a wide range of academic texts. There are three elements associated with this definition that need to be explained: frequency, strength of association and academic register. First, frequency refers to the number of occurrences of academic collocations in a corpus. Frequency has been widely used as an indicator of word usefulness (Coxhead, 2019; Dang & Webb, 2016; Vilkaitė-Lozdiene & Schmitt, 2019; Webb & Nation, 2017). As learners are more likely to encounter high-frequency words, these words should be prioritised in teaching and learning. Nation (2016) suggests that if a word is useful to teach, it will also be useful to test. This is what Laufer and Nation (1999) refer to as a 'cost-benefit' distinction. Time and effort should be better spent on items that will be used frequently. Therefore, this thesis will focus on frequent academic collocations only. Frequency will be used as one of the criteria for selection of collocation items for the ACTs (see Chapter 4, Section 4.3).

Second, strength of association refers to the statistical likelihood that two words co-occur. Measures such as Mutual Information (MI) score, t score or z score (see Gablasova et al., 2017) have been used to measure the strength of word combinations. The MI score will be employed in this thesis to identify academic collocations because this is the most frequently used measure in collocation research (e.g., Fernández & Schmitt, 2015; Nguyen & Webb, 2017; Siyanova & Schmitt, 2008) as well as in corpus tools such as Corpus of Contemporary American English (COCA) (available at https://www.english-corpora.org/coca/). The MI score of 3.00 is a common threshold at which two words appear more frequently than expected by chance and can therefore be called 'a collocation' (Church & Hanks, 1990).

Third, academic register is an important element in the definition of academic collocations because it distinguishes collocations that are significantly more frequent in academic texts from general collocations in non-academic texts. For example, although both *backup plan* and *contingency plan* are collocations with similar meaning, the latter is more frequently found in

academic contexts while the former is often used in less formal situations. This study only focuses on collocations that are useful for academic study in English institutions.

Academic collocations, like general collocations, can fall into two categories: lexical and grammatical. Being aware of this categorisation is necessary to understand why different studies can focus on specific kinds of collocations (see Sections 2.3 and 2.4). Lexical collocations (e.g., *medical treatment* and *background knowledge*) refer to combinations of content words only, including nouns, verbs, adjectives and adverbs. They have been further divided into different kinds based on various combinations of content words. The classification may slightly differ in different studies. For example, Ackermann and Chen (2013) identified eight collocation kinds: 1) adjective + noun, 2) noun + noun, 3) verb + noun, 4) verb + adjective, 5) adverb + adjective, 6) adverb + verb, 7) verb + adverb and 8) adverb + verb participle. Lei and Liu (2018) also include eight lexical collocation kinds, but replace 7) and 8) with noun + verb and adverb + adverb. Lexical collocation categories will be further discussed in Chapter 4 (Section 4.2).

Grammatical collocations (e.g., *based on, this study*), on the other hand, are constructed with a content word and a function word such as prepositions, determiners, conjunctions or pronouns (Durrant, 2009). They have also been classified based on how a content word and a function word are combined, for instance, determiner + noun, conjunction + adverb or adjective + preposition (see Durrant, 2009 for more). With a function word in its component, the meaning of grammatical collocations is sometimes incomplete such as *of income* (preposition + noun) or *no significant* (determiner + adjective). These collocations might be more useful when being expanded to larger multiword units (e.g., *a source of income* or *no significant difference*). The target collocations in this study are lexical or grammatical collocations as complete expressions.

2.2 Why are academic collocations important?

The importance of academic collocations has been highlighted in the English for Academic Purposes (EAP) literature. First of all, this group of lexis facilitates comprehension of academic texts. Following the lexical priming theory (Hoey, 2005), people's minds are primed to automatically link words that they have seen in a chunk before. Therefore, language users tend to process familiar word combinations with less effort than those that they have not already encountered. Furthermore, as academic collocations appear repeatedly in academic texts

regardless of disciplines (Chon & Shin, 2013; Lei & Liu, 2018), knowledge of these words is vital for successful comprehension of course material, including the large number of required and suggested readings in each university course. A lack of academic collocation knowledge may hinder learners' understanding of what being conveyed in academic settings, which in turn may negatively affect their performances in university courses.

In terms of language production, academic collocations are featured in academic writing as markers of proficient language users. At the level of tertiary education, university students are expected to write in a way that is accepted by the academic community. Crossley et al. (2015) argue that collocations are crucial for appropriate language use in academic contexts. These collocations add to the "academic voice" (Vaseghi et al., 2020) which second language (L2) learners need to contribute to knowledge and publish in prestigious journals and establish themselves as credible researchers (p.56). The production of deviant collocations by advanced non-native speakers may indicate a lack of academic expertise (Henriksen, 2013). Taken together, knowledge of academic collocations is an essential component of academic success in English-medium universities.

Now that the importance of academic collocations has been established, the question of which academic collocations should be included in an assessment must be addressed. The following section discusses this issue in relation to corpus-based word lists of academic collocations.

2.3 What are the existing sources of academic collocations for the development of Academic Collocation Tests (ACTs)?

This section begins with a brief overview of existing lists of academic collocations, because a commonly employed procedure when developing vocabulary tests is to make use of word frequency lists to guide the item selection for the assessment (Nation & Webb, 2011). To address the question of how one list might be found to be more suitable for the testing purpose in the current research, methods of evaluating word lists are then reviewed.

2.3.1 Development and validation of academic collocation lists

Several lists of academic collocations have been created with the aim to support pedagogy (e.g., Ackermann & Chen, 2013; Lei & Liu, 2018). Table 2.1 provides a snapshot of four current published lists on academic collocations. These lists will be discussed in the order in which they were published in terms of their features (i.e., number of items, kinds of collocations and item selection), before their validation processes are reviewed.

Table 2.1

Summary of Published Academic Collocation Lists

Word lists	Durrant (2009)	Chon & Shin (2013)	Ackermann & Chen (2013)	Lei & Liu (2018)
Number of items	1,000	460	2,468	9,049
Kinds and examples	Lexical and grammatical E.g., *significant difference, you know*	Lexical and grammatical E.g., *empirical evidence, this research*	Lexical E.g., *widely vary, causal link*	Lexical and grammatical E.g., *achieve goal, based on*
Item selection	Computational extraction	Computational extraction	Computational extraction and manual checking	Computational extraction and manual checking
Validation	No information	No information	Corpus-based comparison	Dictionary comparison

Durrant (2009) employed statistical measures to identify academic collocations in a corpus of 25 million running words of academic written English. His final list for EAP contains 1,000 items, and the majority of them are grammatical collocations. The top five collocations in this list are all grammatical: *this study, associated with, this paper, and respectively,* and *based on*. Durrant (2009) highlights this feature of academic collocations and raises a concern that grammatical collocations may be of little interest to EAP teachers and learners because they "lack the striking salience of collocations" (p.164). Despite its contributions in the field of vocabulary research to demonstrate the nature of academic language, this list might not be suitable for the testing purpose in the current study. This is because the list of Durrant (2009) mostly contains incomplete collocations, whereas the present study focuses on academic collocations with complete meanings (see Section 2.1).

In another study that identified grammatical patterns in academic multiword units, Chon and Shin (2013) developed a list with 460 spoken and 934 written academic collocations, which were all derived from 20 most frequent academic node words (e.g., *obviously, data, process*) in the British Academic Spoken English (Thompson & Nesi, 2001) and the Academic corpus (Coxhead, 2000). These collocations were then classified according to three categories: referent + academic word (e.g., *this analysis*), noun phrases (e.g., *economic development*) and prepositional phrases (e.g., *of income*). It is surprising that adjective + noun and verb + noun collocations were not included in this categorisation, given that they are the two most common kinds reported in research on collocations (e.g., Nesselhauf, 2005; Laufer & Waldman, 2011; Nguyen & Webb, 2017).

Both Durrant's (2009) and Chon and Shin's (2013) lists contain incomplete expressions such as *and respectively* or *of income*. The focus on grammatical collocations results in a rather unstriking group of words. This is important because according to Wulff (2019), the learnability of items can depend on their salience as a result of being unusual. This feature may make the lists receive less attention from EAP teachers and learners than other lists of academic collocations such as the Academic Collocation List (ACL) (Ackermann & Chen, 2013) and the Academic English Collocation List (AECL) (Lei & Liu, 2018) which only include complete expressions (e.g., *significant difference* and *relatively high*). Let us now look at the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018) in turn.

Ackermann and Chen (2013) created the ACL from an academic corpus of 25.6 million words. This list focuses on lexical collocations only. The development process began with corpus-based identification of academic collocations, followed by ranking by an expert panel using a rating scale. Only items selected by the panel as being pedagogically relevant collocations for EAP teachers and learners were included in the final list of 2,469 items. Human intervention makes this study distinct from the other wordlist studies on academic collocations. A similar method is also employed by Simpson-Vlach and Ellis (2010) and Rogers et al. (2021) for their lists of academic multiword units, which are made up of combinations of more than two words. This qualitative judgement, according to Martinez (2019) can be part of the mixed-methods approach to corpus-based research which involves "an interaction among text, potential applications of the research, and the researchers themselves" (p.213).

Developed with the same pedagogical purposes as the ACL (Ackermann & Chen, 2013), the AECL (Lei & Liu, 2018) mostly relied on statistical measures and computer extraction to select 9,049 items for their list. The list was created from an academic written corpus of 43.1 million words, which is the largest corpus of all the studies mentioned above. The AECL includes both grammatical (e.g., *apply to*) and lexical collocations (e.g., *deem necessary*), and Lei and Liu (2018) argue that this list is more balanced than other lists because of this feature. That said, compared to the other three lists of academic collocations, the AECL is huge, which may interfere with its intended pedagogical purpose. Considering limited class time, a list of a more manageable size would better assist EAP teachers and learners to set short-term learning goals (Nation, 2016).

Validation is an important step in wordlist development to evaluate a newly created list and compare it with other pre-existing lists (e.g., see Coxhead, 2000; Dang et al., 2017; Gardner & Davies, 2014). As shown in Table 2.1, however, validation appears to be an optional step in the development of academic collocation lists and it has been carried out in different ways. For example, Lei and Liu (2018) compared their list with a general English collocation dictionary to find out the overlap between these sources. Ackermann and Chen (2013) used a corpus-based approach to compare the overall coverage of the ACL over their academic corpus and its coverage over a general corpus of a similar size. These two approaches both aim to provide evidence that the lists are more academic than general in nature. However, these different approaches make it difficult to compare results across studies. To have a thorough assessment of the lists, it is preferable to combine different approaches to evaluate the lists in the same study. To gain more insight into this issue, let us turn to research on evaluating word lists.

2.3.2 Approaches to evaluating word lists

Research focusing on the evaluation of word lists is relatively rare, but three main methods can be applied: comparing lexical constituents, investigating lexical coverage and using Nation's (2016) evaluation framework. Firstly, the simplest way to examine word lists is to look into lexical constituents between the lists. Lexical analysis can help us understand the extent to which two lists overlap and differ (Hartshorn & Hart, 2016). For example, Hartshorn and Hart (2016) compared the Academic Word List (Coxhead, 2000) and the Academic Vocabulary List

(Gardner & Davies, 2014) and found a 31.05% overlap. Coxhead and Dang (2019) compared three multiword unit lists of spoken four-word sequences in university, namely Biber et al. (2004), Coxhead et al. (2017) and Simpson-Vlach and Ellis (2010) and reported a small overlap among the lists. The advantage of this approach is that it can be easily applied to both single and multiword-unit word lists. However, combining this approach with other evaluating approaches, such as investigating lexical coverage and using Nation's (2016) evaluation framework, may help us gain more insight into the lists.

Secondly, the lexical coverage approach looks at the proportion of words in a text covered by a particular word list (Nation & Waring, 1997). It appears to be the most common criterion used in word list evaluation studies. Coverage data can indicate the extent to which a word list can support users to comprehend a text or raise awareness of the pervasive nature of items from a word list in texts. For example, both Nation (2004) and Dang and Webb (2016) use lexical coverage to compare the General Service List (West, 1953) with other high-frequency word lists, and for EAP, Coxhead (2000) and Gardner and Davies (2014) compared coverage of their lists against an academic corpus and a non-academic corpus. For lists of multiword units, the coverage of Ackermann and Chen's (2013) academic collocation list and Miller's (2020) idiom list was investigated over their academic corpora. That said, there has yet to be a study that compares lexical coverage across lists of multiword units.

Thirdly, Nation (2016) sets up a framework (see Table 2.2) with guiding questions on various features of a list for evaluating word lists. The framework has eight main categories, from setting out the purpose of a word list through to how the list was made, any weaknesses, and whether and how the list has been made available. Nation's (2016) framework appears to be the only available word list evaluation framework currently. Nation (2016) uses his framework to critique his British National Corpus and Corpus of Contemporary American English (BNC/COCA) frequency and supplementary word lists. At first glance, the framework appears quite long and complex, and it seems to be intended for researchers instead of teachers and learners to create or investigate the effectiveness of word lists. Having said that, the framework does cover the main aspects of decisions that list-makers need to employ and report. Some of the guiding questions for evaluation in Table 2.2 are about documentation, for example asking about the target population and purpose of the list. Although these questions are not directly related to the list

quality, they provide important information to evaluate whether two lists can be used in the same context. Nation's (2016) framework appears to be designed specifically for single word lists with details such as unit of counting, homoforms, proper names or acronyms. This framework could be adapted for use with multiword units, and this avenue is explored further in the present study (see Chapter 3, Section 3.1.1).

Table 2.2

Nation's (2016) Framework of Wordlist Evaluation (pp. 131-132)

Focus	Questions
Purpose	Was the target population for the word list clearly described? Was the purpose of the list clearly described?
Unit of counting	Was the unit of counting suited to the purpose? Was the unit of counting clearly defined, including issues such as UK vs US spelling, alternative spellings, part of speech, abbreviations and numbers? Was the unit of counting explicitly well-justified?
Corpus	Was the content of the corpus suited to the purpose of the list? Was the corpus large enough to get reliable results? Was the corpus divided into sub-corpora so range and dispersion could be measured? Were the sub-corpora large enough, of equal size, and coherent? Was the corpus checked for errors?
Main word lists	Was there an explicit description of what would be counted as words and what would not be included? Were homoforms dealt with? Were proper names dealt with, including proper name homoforms? Were content bearing proper names distinguished? Were hyphenated words dealt with? Were transparent compounds dealt with in a way consistent with hyphenated words? Were acronyms dealt with, including acronym homoforms? Were the proper name lists and other lists revised on the basis of initial output?
Other lists	Were marginal words dealt with? Were any other supplementary lists used?
Making the lists	Were the criteria for inclusion and ordering in the list (frequency, range dispersion, or some composite measure) clearly described and justified? Were the criteria for making sub-lists clearly described and justified? Were any subjective criteria used? Were they described and justified? Were the lists checked against competing lists not just for coverage but also for overlapping and non-overlapping words?
Self-criticism	Are the weaknesses of the lists clearly acknowledged?
Availability	Are the lists readily available in electronic form for evaluation?

2.3.3 Summary

This section so far has shown that there are different lists of academic collocations and among them, only the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018) are composed entirely of written academic collocations as complete expressions. As Wulff (2019) suggests, collocations as complete expressions will be more salient to learners and teachers in EAP. Therefore, both the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018) are potential sources of academic collocations for the development of the ACTs in the present study. A question which remains is how these two lists are different or have been evaluated. To the best of my knowledge, no studies have evaluated these lists. The evaluation will provide insights into the similarities and differences between the lists. Consequently, in Chapter 4 of this thesis, the three approaches discussed in Section 2.3.2 are combined to evaluate the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018).

When the question of the best possible source of academic collocations for test development is resolved, another issue that needs to be considered is how knowledge of these words will be assessed (i.e., the task format that will be employed in the assessment). The following section addresses this issue when reviewing previous measures on collocation knowledge.

2.4 How can knowledge of collocations be tested?

This section examines various measures on recognition and recall knowledge of collocations in order to identify any strengths and weaknesses in design, as well as to illustrate how the current study builds on previous research. The collocation tests selected for discussion here (see Table 2.3 for a summary) represent a wide range of test formats which are also among the most frequently cited and updated references in the literature. Although academic collocations and general collocations are not the same, the way they are tested can be similar. Due to the scarcity of studies on academic collocations, measures of general collocations are also discussed here. This section, however, does not cover collocation tests designed for intervention studies because the target collocations are often specific in those testing contexts and do not represent collocation knowledge in general.

Table 2.3

Summary of Previous Measures on Collocation Knowledge

	Study	Target collocation	Test format	Item selection	Validation
Recognition	Gyllstad (2009)	Verb + noun	Multiple-choice (without context, 50 items) Yes/ No (100 items)	Selecting node words from JACET 8000-word list (Ishikawa et al., 2003) and then identifying collocates	Cronbach's alpha Analyses of variances (ANOVAs) Correlation analyses
	Chen (2019)	Verb + noun, Noun + prep, Adjective + noun, Compound noun, Other kinds	Matching (78 items)	Items sampled from a newly created list of 2,000 most frequent collocations	Argument-based approach (Kane, 2006, 2013)
	Wongkhan & Thienthong (2020)	Adjective + noun, Verb + noun, Adverb + verb, Adverb + adjective	Multiple-choice (with context, 10 items)	Collocations from Frankenberg-Garcia (2018)	Not explicitly discussed in the published paper
Recall	Barfield (2009)	Noun-related collocations	Stimulus-response (30 items)	30 target nouns from the 500 most frequent items in BNC; Collocates from Collins Wordbanks Online, Oxford Collocations Dictionary	Not explicitly discussed in the published paper
	Voss (2012)	Verb + noun	Gap-fill (one word, 35 items)	Items sampled from an academic corpus of 16 million running words	Argument-based approach (Kane, 2006, 2013)
	Fernández & Schmitt (2015)	Lexical collocations	Gap-fill (two words, 50 items)	Items varied widely in terms of frequency, t-score and MI score in COCA	Piloting
	Frankenberg-Garcia (2018)	Noun-related collocations	Gap-fill (multi-responses, 10 items)	Nodes: 10 nouns from the AVL (Gardner & Davies, 2014); Collocates: in PICAE	Not explicitly discussed in the published paper

2.4.1 Measuring recognition knowledge of collocations

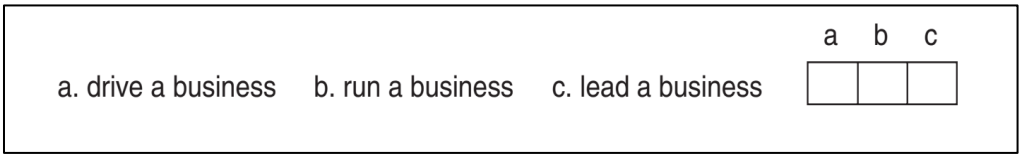
This section discusses three studies of Gyllstad (2009), Chen (2019) and Wongkhan and Thienthong (2020). The first two measure general collocation knowledge, whereas the other assesses academic collocation knowledge. These studies differ in terms of test format, item selection and validation evidence (Table 2.3), as elaborated below.

2.4.1.1 Gyllstad (2009)

Gyllstad (2009) devised two measures to assess recognition knowledge of verb + noun collocations: COLLEX (50 items) and COLLMATCH (100 items). COLLEX employs the multiple-choice format in which test-takers are required to select one of three options that is the most frequent collocation (see Figure 2.1). This format is practical because it is easy to complete and score, and it is also familiar to learners from a variety of language backgrounds. However, items in COLLEX are context-independent, which lessens the authenticity of the test task and makes it less representative of similar tasks in academic settings. The usefulness of the test scores is also reduced because it cannot be inferred from the test results whether test-takers understand the meaning of the target collocations. Another variation of this multiple-choice format with context sentences will be discussed in Section 2.4.1.3.

Figure 2.1

Example of a COLLEX Item (Gyllstad, 2009, p.157)

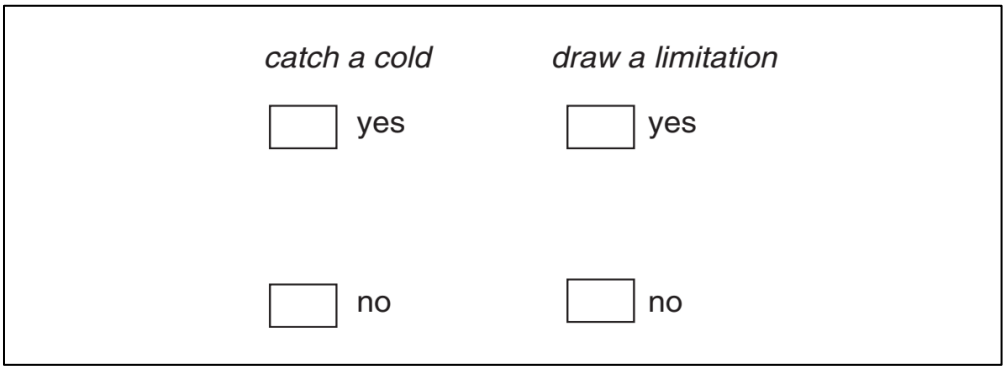


COLLMATCH uses a Yes/ No format and requires test-takers to determine whether the presented word combinations occur frequently in English or not (Figure 2.2). This format has the advantages of being easy to create and allowing the assessment of a large number of test items in a short period of time. However, interpreting the test scores can be challenging because test-takers may overestimate (i.e., choose Yes when they do not truly understand the words) or underestimate their knowledge (i.e., choose No when they actually know the words). Martinez (2011) expresses concern that “there is no

direct check on task-taker understanding of the words tested, but rather relies on the 'honesty' of the person being tested" (p.146). Lee and Shin (2021) compared different test formats for measuring collocation knowledge and stated that the Yes/No task is much easier and less discriminating compared to the multiple-choice format. Thus, the Yes/ No format has not been employed for the AC Recognition Test in the current study.

Figure 2.2

Example of a COLLMATCH Item (Gyllstad, 2009, p.158)



For item selection, Gyllstad (2009) used the JACET 8000-word list based on British National Corpus (BNC) (Ishikawa et al., 2003) and selected node words from the first 5,000 word levels for COLLEX and from the first 4,000 word levels for COLLMATCH. The collocates of those node words were then identified in the BNC. While items for COLLEX were selected based on corpus statistical measurement (i.e., frequency and z-scores), those for COLLMATCH relied on the intuition of the author and two lecturers of English. As Gyllstad (2009) acknowledged, the item sampling following the word-centred approach, in which single words were selected as a starting point, could not represent the entire population of high-frequency collocations. It is also unclear why the JACET 8000-word list (Ishikawa et al., 2003) was selected for test development, as well as why different criteria for item selection were applied for COLLEX and COLLMATCH, even though both tests focus on recognition knowledge of verb + noun collocations.

For validation evidence, Gyllstad (2009) administered COLLEX, COLLMATCH and a modified version of the Vocabulary Levels Test (VLT) (Schmitt et al., 2001) to 307 participants, including 273 ESL learners at four different proficiency levels and 34 native speakers. The findings showed that

both COLLEX and COLLMATCH were reliable in terms of internal consistency (Cronbach's alpha > .89). Gyllstad (2009) concluded that the tests had an acceptable discriminatory power, although the differences between groups were not always statistically significant. Gyllstad (2009) also compared results of COLLEX, COLLMATCH and the modified VLT. The results indicated a high correlation between COLLEX and COLLMATCH scores, and they both were strongly correlated with the modified VLT scores (r > .80). Although these findings contributed to a better understanding of learners' collocation knowledge, the validation process of the tests was not guided by any coherent framework, making it difficult to interpret how the research findings were related to test validity.

2.4.1.2 Chen (2019)

To measure recognition knowledge of collocations, Chen (2019) adopted the matching format used in the VLT (Schmitt et al., 2001) to test five groups of general English collocations (Table 2.3) with a total of 78 test items (26 clusters with three target collocations each). Sample test items are presented in Figure 2.3. Chen (2019) also created a list of the 2,000 most frequent collocations for sampling, as well as a C-test of general English proficiency and a multiple-choice test of routine formulae (i.e., pragmatic expressions) for the purpose of validation. The validation process followed the argument-based approach (Kane, 2006, 2013) and consisted of two phases. Phase One had 179 participants and Phase Two had 397. Chen's (2019) findings suggested that collocation test scores significantly correlated with both C-test and routines scores. Collocation kind, frequency, degree of coherence and semantic transparency were also found to be significantly correlated with test item difficulty.

Figure 2.3

Example of Test Items With Answers in Chen (2019, p.531)

provide (…)____	a	service
[A: a]	b	minute
reach (…)____	c	time
[A: f]	d	power
wait (…)____	e	responsibility
[A: b]	f	agreement

Chen's (2019) study has some merits in terms of considering multiple aspects of test design. The inclusion of various collocation kinds in testing better represents target language use in real life. Moreover, the validation process was guided by a coherent framework, which is currently the mainstream in the language testing field (see more in Section 2.5.2). Overall, Chen (2019) is an example of an ambitious study which included the development of a collocation list and three newly designed tests (i.e., the collocation test, the general proficiency test and the routines test) in a single project. As Chen's (2019) collocation list and collocation test are more relevant to the current study, they are discussed more closely as follows.

Chen's (2019) collocation list was derived from a purchased list of the most common two-word collocations for the top 5,000 lemmas from COCA (Davies, 2008-2016). The size of the original list was unknown, but after Chen (2019) applied some criteria (e.g., nodes and collocates from the 2,000 most frequent lemmas and MI score ≥ 3) to remove items, the final list contained the 2,000 most frequent collocations in the corpus. It is important to note that this list used for test item sampling includes items such as *black hair, blue eye,* or *buy house* which are free combinations rather than collocations. Another important note is that this list needs to be evaluated to ensure that the items included are representative of those in the target language use domain. Unfortunately, Chen (2019) did not use any of the evaluating methods mentioned in Section 2.3.2 to evaluate this newly created list.

With respect to the collocation test, Chen (2019) did not explicitly specify the sampling rate or justify how the number of test items could adequately represent the population of frequent collocations. Additionally, the matching format used by Chen (2019) is not a popular choice for measuring recognition knowledge of collocation. When designing their multiple-choice collocation test, Nguyen and Webb (2017) also considered the matching format but rejected it due to the interdependence of the test items. The main issue is that the correct answer to one test item serves as a distractor for other items in the same cluster, which does not always work well. Taking the cluster in Figure 2.3 as an example, *minute* is the correct collocate of *wait,* and *time* can be a good distractor, but not *service* or *agreement*. Moreover, a mismatch can leave an undesirable trace on test-takers' memories, which potentially affects their future language use (see Boers, 2021; Boers et al., 2014). Furthermore, Chen (2019), like Gyllstad (2009), employed context-independent test items, which differs from how the

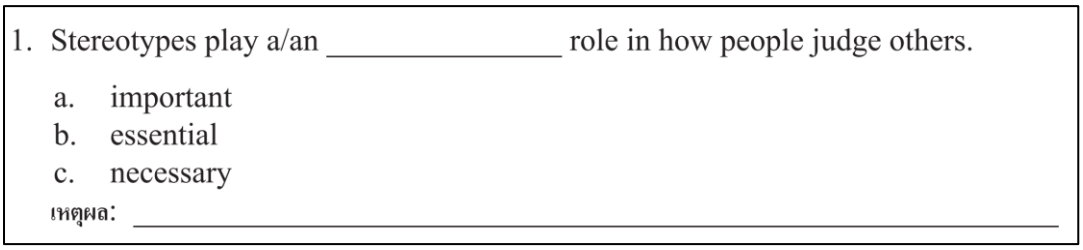
collocations are used in real-life settings. To address this gap, the following section will look at a study that makes use of context sentences to design a collocation test.

2.4.1.3 Wongkhan and Thienthong (2020)

Wongkhan and Thienthong (2020) created a ten-item multiple-choice recognition measure of academic collocation knowledge. Their test contains context sentences without target collocates. Test-takers are asked to fill in the gap by choosing one of three synonyms as collocate options, and they have to justify their selection on the space provided (see Figure 2.4 for a sample test item). After administering the test to 120 Thai undergraduate students, Wongkhan and Thienthong (2020) found that learners tended to select the collocations with high frequency. Their findings also indicated that students with more academic experience (i.e., education years) had better knowledge of academic collocations than their less experienced counterparts.

Figure 2.4

Example of Test Items in Wongkhan and Thienthong (2020, p.14)


1. Stereotypes play a/an _____ role in how people judge others.

   a. important
   b. essential
   c. necessary

   เหตุผล: _____

Wongkhan and Thienthong's (2020) study suffers from several limitations. Firstly, when context sentences are employed, it is important to ensure that the elicited knowledge of collocations is not hindered by comprehension of the sentences (Read, 2000). Taking a test item in Wongkhan and Thienthong (2020) as an example (Figure 2.4), "stereotypes" is used in the sentence prompt. It belongs to the 4,000 word level and is therefore a mid-frequency word according to Nation's BNC/COCA lists (2012). This word might not have been familiar for test-takers because it is not a high-frequency word. Focusing too much on the unknown word may distract test-takers from choosing the correct answer. Secondly, there was a lack of justifications for item selection in the study by Wongkhan and Thienthong (2020). One of the core issues in test development is determining the number of test items and justifying how the sampling rate adequately represents the

knowledge of collocations (Gyllstad & Schmitt, 2018). With only ten items in the test by Wongkhan and Thienthong (2020), generalisations from the test results are unable to be drawn. Finally, this test was published without any discussion of the validity evidence, which is critical for test users to evaluate test quality and decide whether or not to use it.

So far, the review has highlighted several issues related to test design and validation of previous tests on recognition knowledge of collocations. The next section turns to measures of recall knowledge of collocations before all the assessment issues are summarised in Section 2.4.3.

2.4.2 Measuring recall knowledge of collocations

The four main studies reviewed in this section are representative of various studies on measuring recall knowledge of collocations. As shown in Table 2.3, these collocation tests are varied in terms of the target collocations, test format, item selection and validation. It should be noted that although translation has been used as an assessment method of collocation knowledge (e.g., Bahns & Eldaw, 1993; Gitsaki, 1999; Macis & Schmitt, 2017), it is not included in this review because the ACTs in the present study are designed for learners with various language backgrounds, while the translation task is more appropriate for groups of learners with the same first language (L1).

2.4.2.1 Barfield (2009)

Barfield (2009) developed a 30-item stimulus-response collocation test called LexCombi to measure recall knowledge of 89 Japanese university students from a wide range of proficiency. In this test, 30 stimulus nouns were selected from a list of the 500 most frequent items in the BNC. The database of possible collocates was built from Collins Wordbanks Online and the Oxford Collocations Dictionary, which was then used for scoring learners' responses on LexCombi. Test-takers were asked to provide three collocates for each noun. Figure 2.5 shows an example of a test item from LexCombi with accepted collocates such as *hard, at* and *voluntary*. The results suggested that advanced learners outweighed their less proficient counterparts both in the quantity and the appropriateness of the collocates produced. Although Barfield (2009) employed different analytical methods such as Rasch model for item analysis and t-test for comparisons between groups, these findings were not discussed in relation to the validity of the test.

Figure 2.5

Example of a Lexcombi Item (Barfield, 2009)

work	1._____	2. _____	3. _____

A positive feature of LexCombi is that it can elicit a considerable number of collocations from learners in a quick and effective way. However, it is not clear how a collocation is defined in Barfield (2009). For example, according to Barfield's (2009, p.99) database, there are a total of 113 possible collocates that can combine with the noun "work", many of which constitute a free combination instead of a collocation (e.g., *his, her, this*). The format of LexCombi is less appropriate to adopt for the present study because the respondents might produce collocations which are not the ones being tested in the recognition version. A more restricted-response test format is, therefore, more suitable for the current research, such as those in Voss (2012) and Fernández and Schmitt (2015).

2.4.2.2 Voss (2012)

Voss (2012) developed a Collocational Ability Test to assess 206 ESL learners' recall knowledge of 35 verb + noun collocations. For each test item, test-takers were provided with an English sentence containing a target collocation and they had to fill in a gap with one word of the collocation pair. The target collocations were sampled from the written academic sub-corpus of the BNC (16 million running words). The context sentences, which helped to infer the meaning of the collocations, were selected from concordance lines in the academic corpus. Figure 2.6 shows an example of a test item in Voss (2012) for the collocation *exert influence*. In addition to the Collocational Ability Test, Voss (2012) also used a self-created 30-item vocabulary size test based on Laufer and Nation's (1999) productive levels test and a reading sub-test from the Michigan English Language Assessment Battery (MELAB).

The results showed that (a) the ESL learners had limited recall knowledge of academic collocations with an average score of less than 20%; (b) knowledge of academic collocations moderately correlated with vocabulary size; (c) there was a moderate relationship between knowledge of academic collocations and reading skill; and (d) depending on the scoring method, the relationship

between test item difficulty and frequency of academic collocations was non-significant (dichotomous scoring) or moderate (polytomous scoring). With the dichotomous scoring, a response was marked either correct or incorrect, while with the polytomous scoring method, test-takers received a partial score if their answer was partially correct. It should be added that the study of Voss (2012) was oriented as a validation study of the Collocational Ability Test instead of vocabulary research on knowledge of academic collocations. Therefore, the findings served as validity evidence in Voss's (2012) research rather than an explicit discussion on the development of academic collocation knowledge.

Figure 2.6

Example of a Collocational Ability Test Item (Voss, 2012, p.210)

Although it might be very easy to steal from a friend or colleague and not get caught most people would feel this to be "wrong", yet such feelings may not _____ such a strong influence over decisions as to whether to steal from a larger and less personal victim.

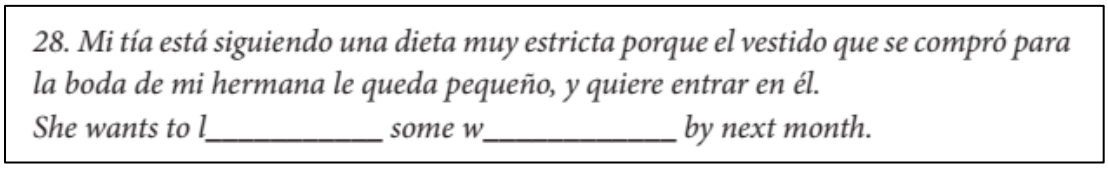
The study of Voss (2012) has both merits and limitations in its design. Collocations in Voss (2012) were not selected based on the frequency of individual words as they were in Gyllstad (2009) or Barfield (2009), but on the frequency of the whole units as they occurred in an academic corpus. This approach to the identification of collocations, therefore, better represents target language use. The use of concordance lines as context sentences for test items also helps to increase ecological validity. That said, the authentic academic contexts can contain long sentences, infrequent words and complicated grammar (as shown in Figure 2.6), which may lead to learners giving incorrect answers and might in turn affect construct validity. Moreover, the format seems to encourage test-takers to think of the collocations as separate words, not as a unit (see Boers et al., 2014).

2.4.2.3 Fernández and Schmitt (2015)

Also employing the gap-filling format like Voss (2012), Fernández and Schmitt (2015) used initial letter hints and presented a collocation as the whole unit in their test. Test-takers are asked to provide two-word collocations gapped in an English sentence which summarises the context in their L1 (i.e., Spanish). The first letter of each word of a collocation pair is given. In the example test item in Figure 2.7, the Spanish context means "*My aunt is following a very strict diet because the dress that was bought for my sister's wedding is small, and she wants to wear it*", and the answer is "lose weight" (Fernández & Schmitt 2015, p.103). The 50-item collocation test was administered to 108 Spanish speakers from a wide range of proficiency levels. The findings indicated that a) the participants could recall a considerable number of collocations; b) the frequency of a collocation as a whole was more important factor than MI score or t-score for L2 learning of collocations, demonstrated by a positive relationship between raw corpus frequency with collocation knowledge (r = .45); c) the correlation between years of study and knowledge of collocation was also moderate (r = .45).

Figure 2.7

Example of a Test Item in Fernández and Schmitt (2015, p.103)



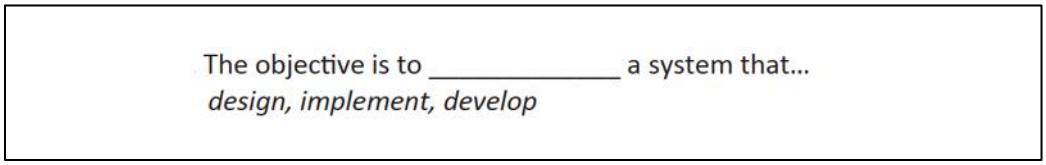
The test format in Fernández and Schmitt (2015) was very directive, and it could narrow down the possible options for test-takers. For example, with the item in Figure 2.7, test-takers could easily figure out that *slim down* is not an option that can fit, because the target words start with *l* and *w*. The format involves the use of L1, which means the test may not be suitable for testing groups of participants of various language backgrounds. That said, a similar format can be adapted for the purpose of the present study by replacing the L1 content with sentences in English.

2.4.2.4 Frankenberg-Garcia (2018)

The recall test in Frankenberg-Garcia (2018) employs another variation of the gap-filling format. Test-takers are provided with ten gapped sentence excerpts based on collocations for ten nouns selected from the Academic Vocabulary List (AVL) (Gardner & Davies, 2014) and are asked to fill the gaps with as many collocates as possible for the given contexts (see Figure 2.8). The test was administered to 90 EAP writers (students and staff members) in a British university, including both native English speakers (English L1) and speakers of other languages (Other L1) who had very high levels of English proficiency. The findings suggested that the difference in the use of academic collocations was not significant between English L1 writers and Other L1 counterparts, and the number of collocations produced correlated with academic experience, as indicated by participant role in higher education: undergraduates the least experienced group and the academics as the most experienced.

Figure 2.8

Example of a Test Item in Frankenberg-Garcia (2018, p.97) With Answers in Italics



Frankenberg-Garcia (2018) provides useful insights into a range of academic collocation choices that EAP writers are able to recall about the ten target nouns from the AVL. Nevertheless, test-takers are required to fill in only one word which may navigate their thinking towards individual words instead of collocations (see also Voss, 2012). For the purpose of the present study to compare recognition and recall knowledge of the same target collocations, this test format is not appropriate as too many possible answers can fit the given contexts.

2.4.3 Summary

Several lessons can be learnt from previous measures of collocation knowledge for the purpose of developing the ACTs in the present study. With respect to the test formats, the multiple-choice for the recognition test and the gap-filling for the recall test are the two most commonly used in previous

collocation tests and are the best candidates for the ACTs in the present study. The main issue with these test formats in prior tests is that only one word of a collocation pair is tested, which might encourage test-takers to think of individual words instead of collocations (e.g., Frankenberg-Garcia, 2018; Voss, 2012; Wongkhan & Thienthong, 2020). Another issue is the use of context-independent items which is different from how the collocations appear in real-life settings (e.g., Gyllstad, 2009). Consequently, to better replicate the target language use domain, the target collocations in the ACTs are embedded in context sentences and tested as whole units.

In terms of item selection, to overcome the issue of item representativeness in previous tests, the ACTs in this study have been developed from the list of Ackermann and Chen (2013) based on results of a wordlist evaluation process (see Chapter 4). By doing this, the ACTs could include various collocation kinds in the list rather than just one kind such as verb + noun collocations, as in Gyllstad (2009) or Voss (2012). The item representativeness could also be enhanced because the ACT items reflect a finite list instead of an entire corpus. In other words, the ACTs can avoid the issue of using a limited number of test items to represent a vast number of items in a corpus, such as ten items in Wongkhan and Thienthong (2020) or 35 items in Voss (2012) representing overall academic collocation knowledge.

Turning to validation process, this study follows Voss (2012) and Chen (2019) in applying the argument-based validation framework (Kane, 2004, 2013) to the ACTs. Unlike other collocation tests which have been published without validation evidence (e.g., Frankenberg-Garcia, 2018), or without the use of coherent framework (e.g., Gyllstad, 2009), Voss (2012) and Chen (2019) present and connect all their empirical findings in a comprehensive framework to evaluate the validity of the interpretations and uses of the test scores. As the argument-based validation is currently the mainstream in the language testing field for its systematic approach to investigate the validity of a language test (Chapelle & Lee, 2021), this study aims to follow this approach and advocate the field of vocabulary assessment with its application to the ACTs.

It is likely that most of the above-discussed issues of previous collocation tests originate from test development and validation processes not being guided by theoretical frameworks that can form the foundation for the test design and the empirical research into the test validity. The following section will look into such frameworks more closely.

2.5 What frameworks can be applied for the development and validation of the ACTs?

Two frameworks form the basis for the ACTs in the present study: evidence-centred design (Mislevy, 2007; Mislevy & Yin, 2013) for test development and argument-based validation (Kane, 1992, 2004, 2013) for validity assessment. These frameworks have been used in conjunction as a package for test development and validation (Fulcher & Davidson, 2007). The close connection between these two is shown in the requirement for evidence to support the design decisions and the claims made about the test scores. These frameworks are introduced in more detail in this section.

2.5.1 Evidence-Centred Design (ECD) for the ACTs

Evidence-centred design is "a conceptual framework for the design and delivery of educational assessments, organised around the idea of assessment as evidentiary argument" (Mislevy & Yin, 2013, p.208). The idea behind the ECD is to help test designers carefully think about the knowledge to be measured and the expected performance of test-takers in a specific context. The full complexity of the ECD can be found in a series of publications of Mislevy and colleagues (e.g., Mislevy et al., 2003; Mislevy et al., 2004; Mislevy & Haertel, 2006). Following Mislevy and Yin (2013), the ECD basically consists of five test design layers:

- **Domain analysis**: In this layer, information about the target construct is gathered to ensure that the tasks in the assessment accurately reflect the target language use domain (TLU). In the case of the ACTs, the TLU is an English-medium academic setting. The domain analysis of the ACTs can begin with an evaluation of corpus-based word lists on academic collocations to select the source of items that are more representative of the academic domain (see Chapter 4). This layer also includes the analysis of literature on previous measures of collocation knowledge in order to identify suitable tasks for test-takers to demonstrate knowledge of academic collocations (as presented in Section 2.3).
- **Domain modelling**: This layer describes the design pattern of the assessment as well as the characteristics of test-takers' performance to constitute evidence of their knowledge. The domain modelling, therefore, includes the description of the item sampling, the characteristics of the test tasks and the expected performances of test-takers. Together with the domain

analysis, the domain modelling served as the foundation for the development of the ACTs in Chapter 4.

- **Conceptual assessment framework**: In this layer, technical aspects of the assessment are considered, including specifications for test tasks, measurement models for analysing statistical characteristics of test items and scoring procedures.
- **Assessment implementation**: This layer concerns the practical elements before the test administration. These include activities such as fine-tuning test items, piloting items and scoring. The development of the ACTs in this study also includes these procedures (see Chapter 4).
- **Assessment delivery**: This layer is related to the process of administering the test. The focus of this layer is on actual interactions between test-takers and test tasks (i.e., how the test is displayed to participants). How test data are recorded and how test scores are reported to stakeholders are also among the considerations.

The higher the stakes of a language test, the more care and rigor must be applied to its design. The TOEFL iBT (Chapelle et al., 2008) is an example of a high-stake test that employed the ECD to guide test development. The ACTs in this study are intended to be low-stakes tests for diagnostic purposes. This means they incorporate some aspects of the ECD rather than fully implementing the entire framework in all its complexity. This thesis explicitly focuses on three layers of the ECD which guide the development process of the ACTs (see Chapter 3): domain analysis, domain modelling and assessment implementation. The two remaining layers – conceptual assessment framework and assessment delivery – are implicit in the analysis of test item characteristics in Chapter 5 and the test administration in Chapter 3. Now that the framework for the development of the ACTs has been presented, the next section will look at the framework for the validation process.

2.5.2 Argument-based validation for the ACTs

The argument-based validation was developed primarily by Kane (1992) to offer a pragmatic approach for measurements in education and psychology. Through a series of publications by Kane (2001, 2004, 2006, 2011, 2013), this framework has become influential in language testing and has been further promoted by other researchers such as Bachman (2005), Bachman and Palmer (2010), Chapelle et al. (2008) and Chapelle (2012, 2020). The logic behind the argument-based approach is

that validation is a process of making claims about the interpretations and uses of test scores and gathering evidence to evaluate these claims. This evidence can be collected through quantitative and qualitative methods such as correlation analyses or retrospective interviews of test-taking processes (Chapelle, 2020). The two main components of this validation framework are an interpretive argument and a validity argument, as explained in more detail below.

2.5.2.1 Interpretive argument

An interpretive argument is an explicit statement of proposed interpretations and uses of test scores. It consists of a sequence of inferences leading from observations of test-takers' performances to conclusions and decisions based on those performances. The number of inferences depends on what is being claimed about the test scores. For example, Kane et al. (1999) illustrate an interpretive argument with three inferences (Evaluation, Generalisation and Extrapolation), while Chapelle et al. (2008) employ six inferences (Domain description, Evaluation, Generalisation, Explanation, Extrapolation and Utilisation). Below is a brief explanation of each inference.

- **Domain definition**: This reference links test-takers' performance to the performance in the target language use domain.
- **Evaluation**: This inference links an observed performance elicited by test tasks to an observed score. In other words, the interpretation of test results starts from the assessment of test characteristics, testing conditions and scoring procedures to produce an observed score.
- **Generalisation**: This inference extends the interpretations to the expected performance over different occasions, different testing contexts and different raters.
- **Extrapolation**: This inference connects the interpretation of test scores to the target domain. Kane (2004) divides the Extrapolation inference into two types:
    - ***(1)*** Extrapolation from test scores to constructs as defined by the theory that the test is based on. For example, scores on a test which is attributed to a construct of academic collocation knowledge should relate to the performance on a test that measures academic vocabulary, as expected theoretically. This first type of Extrapolation inference is also known as *Theory-based inference* (Kane, 2006, 2013).

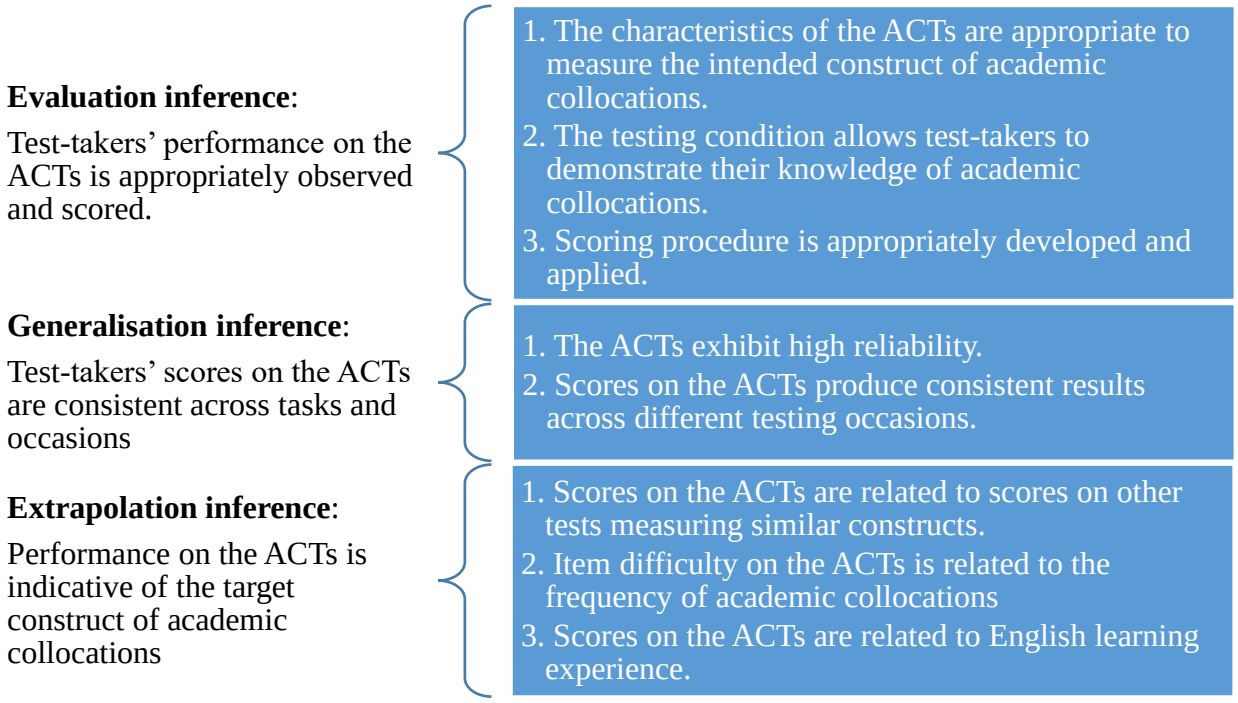
(2) Extrapolation from test constructs to specific activities in practice or real-life situations. For example, scores on an academic collocation test should be related to academic reading or writing skills in a real academic context.

- **Explanation**: Chapelle et al. (2008) make a clear distinction between the two types of Extrapolation inference in Kane's (2004) model by adding the Explanation inference to link test scores to test construct. This inference is similar to the first type of Extrapolation inference in Kane (2004) mentioned above.
- **Utilisation**: This inference links test scores to decisions such as assigning courses or identifying current levels of performance.

This study follows Kane et al. (1999) to include three inferences in the interpretive argument for the ACTs. These inferences are based on the warrants (i.e., general statements that are used to justify a claim), as summarised in Figure 2.9 below.

Figure 2.9

Summary of Inferences and Warrants of the ACTs



2.5.2.2 Validity argument

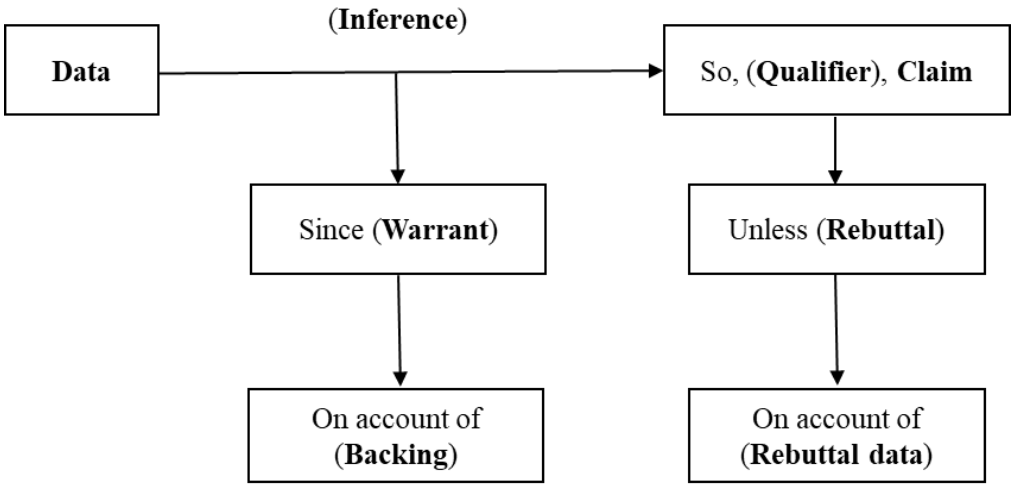
A validity argument gives a critical evaluation of the inferences stated in the interpretive argument in terms of completeness, coherence and degree of support. To successfully construct the validity argument, it is necessary to understand the following key concepts:

- *Data* are performances on a measurement or test scores.
- An *inference* is the link from the data to a claim.
- A *claim* is a proposed interpretation and use based on the data.
- A *warrant* is a general statement that is used to justify a claim.
- *Backing* is the evidence collected from theoretical and/or empirical research to support the warrants.
- A *rebuttal* is an alternative explanation or a counterargument that rejects a claim.
- *Rebuttal data* refer to the evidence that either supports or rejects the rebuttal.
- A *qualifier* shows the degree of confidence in a claim. The qualifier can be qualitative (e.g., sometimes, almost, or always) or quantitative (e.g., standard errors of measurement).

All of the above-mentioned elements are connected in an argument structure as shown in Figure 2.10.

Figure 2.10

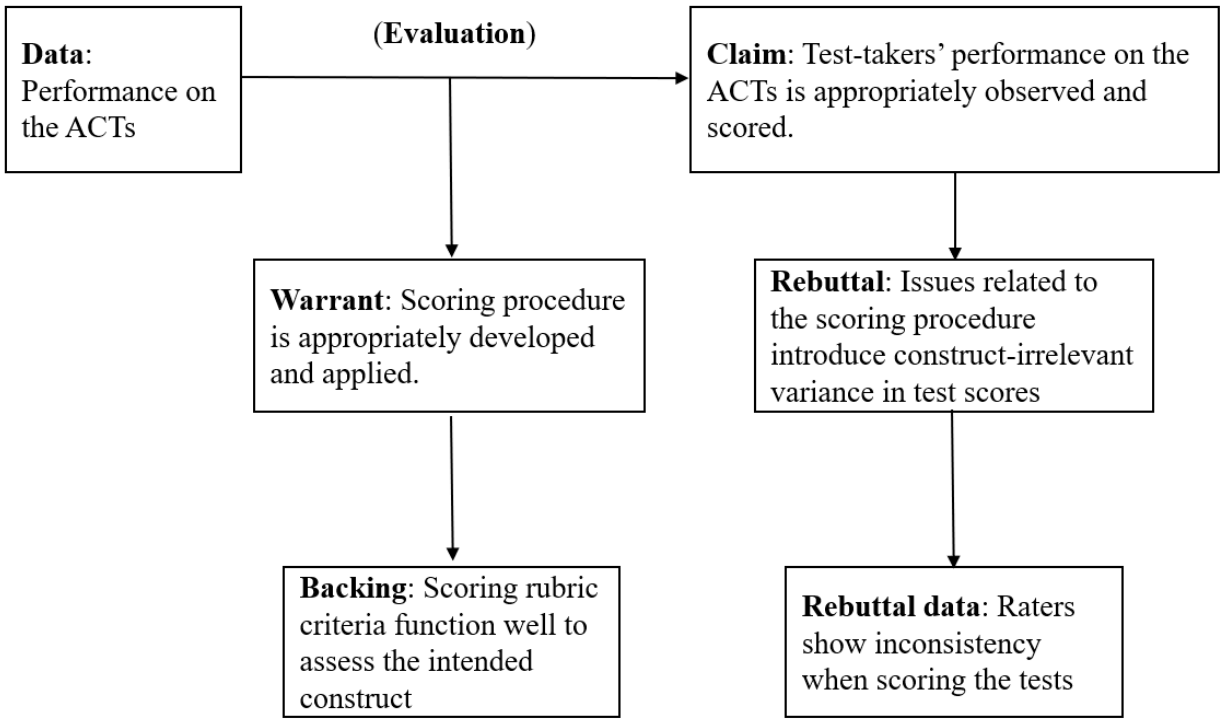
The Argument Structure (Based on Kane, 1992 and Toulmin, 2003)



Let us take the Evaluation inference of the ACTs as an example. Based on test-takers' performance on the ACTs (**data**), the **claim** made for the Evaluation **inference** is that "test-takers' performance on the ACTs is appropriately (**qualifier**) observed and scored". There are several **warrants** to support this claim. One of them is that "scoring procedure is appropriately developed and applied.". **Backing** for this warrant includes scoring rubric criteria which function well to assess the intended construct. One of the **rebuttals** is that "issues related to the scoring procedure introduce construct-irrelevant variance in test scores". **Rebuttal data** can include the fact that raters show inconsistency when scoring the tests. Overall, if the evidence collected supports the warrant and no evidence is found for the rebuttal, the Evaluation inference will stand then. Otherwise, the Evaluation inference will be weakened (with rebuttal data) or rejected (without backing). A visualisation of this inference is provided in Figure 2.11.

Figure 2.11

Example of the Evaluation Inference of the ACTs



An inference can be supported to a certain extent or rejected depending on the evidence collected from empirical research. Different inferences require different kinds of evaluation. For the Evaluation

inference, evidence can be collected from analysis of test-taking conditions, test characteristics and scoring procedures (Dursun & Li, 2021). Quantitative evidence, such as using a Rasch model (Rasch, 1993) to analyse statistical characteristics of test items, is commonly used in validation research (e.g., Brown, 2018; Chen, 2019; Voss, 2012). Qualitative evidence is also employed in several studies for the assessment of the Evaluation inference. Voss (2012), for instance, conducted a follow-up interview with six out of 206 test-takers to investigate their meta-cognitive strategies during the Collocational Ability Test, including test-takers' awareness of the test purpose, their understanding of academic language, their thinking process about academic language when taking the test, possible difficulty with reading comprehension, and information about their English learning experience. These qualitative data are useful because according to Cohen (2006), looking into test-takers' behaviours, including the kinds of strategies that they employ to produce answers, can aid in the validation of test-taker engagement of test tasks.

Concerning evidence for the Generalisation inference, reliability indices are frequently utilised to indicate the consistency of test scores. For example, Chung (2014) reported Cronbach's alpha for a test of productive English grammatical ability in academic writing, whereas Tran (2020) used Rasch reliability estimates for a locally created listening test as backings for the Generalisation inference. As for the Extrapolation inference, correlation analysis is a popular method of gathering evidence to demonstrate that test scores appropriately reflect the intended test construct, especially when a new test is developed (Dursun & Li, 2021). An example of this is the study carried out by Youn (2013) in which the high correlation between a newly created L2 pragmatic test and TOEFL speaking tasks gives support for the validity of the pragmatic test. In a similar case, Oh (2018), when developing a new writing test, used Oxford Grammar Test (Oxford University Press, n.d.) as a criterion test, and a high correlation between the two tests indicates that the construct of the writing test is associated with other criteria of language proficiency. The methods used for collecting empirical evidence in previous validation studies provide useful references for the process of gathering validity evidence for the ACTs in the present study. In the next section, the specific research questions whose answers provide the evidence for the assessment of the inferences in the validation framework of the ACTs will be presented.

2.6 Research questions

The following six research questions, which are connected to the interpretive argument for the ACTs (Figure 2.9), have guided the empirical research in this thesis:

RQ1. To what extent are the characteristics of the ACTs appropriate to measure the intended construct of academic collocations? (Evaluation inference)
RQ2. Does the testing condition allow test-takers to demonstrate their knowledge of academic collocations? (Evaluation inference)
RQ3. Are scores on the ACTs reliable? (Generalisation inference)
RQ4. Are scores on the ACTs related to scores on other tests measuring similar constructs? (Extrapolation inference)
RQ5. Is the item difficulty on the ACTs related to the frequency of academic collocations? (Extrapolation inference)
RQ6. Are scores on the ACTs related to English learning experience? (Extrapolation inference)

The evidence collected for these research questions provides necessary support (partial or complete) to warrant the inferences in the interpretive argument of the ACTs.

2.7 Chapter summary

This chapter started by providing a definition of academic collocations and explaining their importance. Next, it discussed the possibility of using published word lists for the development of the ACTs, including what the existing word lists are and what approaches can be used for evaluating the lists to choose the best source of collocations for test items. The chapter then reviewed previous collocation tests to learn from their strengths and weaknesses. The chapter presented the frameworks which lay the foundation for a robust test development (evidence-centred design) and a comprehensive validation process (argument-based approach) of the ACTs. Finally, Chapter 2 ended with the research questions that guided the empirical study to collect the validity evidence for the ACTs. The next chapter presents the methodology employed for the present study.

Chapter 3 Methodology

This chapter presents the methodology which guided the development of the Academic Collocation Tests (ACTs) and the collection of empirical evidence for the validation purpose. Chapter 3 first describes the steps involved in the wordlist-based development of the ACTs within the Evidence-Centred Design (ECD) framework (Mislevy & Yin, 2013) in Section 3.1. Next, Section 3.2 explains the validation process which was led by the argument-based validation framework (Kane, 2013). This section provides detailed information about the participants, the materials and instruments employed in this study, and the procedures for data collection and data analysis. Finally, Section 3.3 summarises this chapter and gives the rationale for the next chapter.

3.1 Development process of the ACTs

This section outlines six steps involved in the creation of the ACTs within the ECD framework (see Table 3.1). Chapter 4 will go over each step in detail. This section mainly focuses on the methodology for wordlist evaluation in Step 2, as well as participants and procedures for piloting stage in Step 6.

Table 3.1

Development Framework of the ACTs

Wordlist-based development of the ACTs	Evidence-centred design layer
Step 1. Identify available word lists of academic collocations Step 2. Evaluate academic collocation lists	Domain analysis
Step 3. Sample academic collocation items from the selected word list Step 4. Select test formats Step 5. Write test items	Domain modelling
Step 6. Pilot and finalise the ACTs	Assessment implementation

3.1.1 Domain analysis

The first two steps in Table 3.1 belong to the domain analysis layer of the ECD framework. They aim to identify the most suitable source list from which representative items of academic collocations can be selected. As previously discussed in Chapter 2 (Section 2.3), the Academic Collocation List (ACL) (Ackermann & Chen (2013) and the Academic English Collocation List (AECL) (Lei & Liu, 2018) are two potential sources of items for developing the ACTs. In order to provide a thorough assessment of these lists, this study combined three different methods. Nation's (2016) evaluation framework first provided the overall picture of how the lists were developed and validated. Next, the lexical constituents comparison allowed a closer look at how the academic collocations in the lists are similar and different. Lexical coverage then helped to further investigate the academic nature of the collocations in the lists. These three evaluation methods are now described in more detail in turn.

First, Nation (2016) is currently the only available framework for evaluating word lists. However, as this framework is designed specifically for individual words, it needs adapting to be used with lists of multiword units. Several modifications (see Table 3.2) were made to Nation's (2016) original framework (see Table 2.2, Chapter 2) as follows:

- "Purpose" and "Target audience" in the original framework are separated into two categories for greater clarity.
- A new category called "Number of items" has been added for instant capture of the list size.
- "Making the lists" in the original framework is broken down into four categories of "Word selection principles", "Manual checking", "Ordering items" and "Validation" to gain a deeper insight into the list development process.
- The categories "Main word lists" and "Other lists" (e.g., marginal words) in Nation's (2016) framework are not relevant to lists of multiword units and are thus omitted.

Table 3.2

Framework for Evaluating Multiword Unit Lists (Adapted From Nation, 2016, pp. 131-132)

Focus	Questions
A. **Target audience**	Was the target population for the word list clearly described?
B. Purpose	Was the purpose of the list clearly described?
C. **Number of items**	Was the number of items manageable for the intended purpose of the list?
D. Unit of counting	Was the unit of counting clearly defined? Was the unit of counting explicitly well-justified? Was the unit of counting consistently applied?
E. Corpus	Was the corpus size large enough to get reliable results? Were the corpus text types suited to the purpose of the list? Was the corpus divided into sub-corpora? Were the sub-corpora large enough, of equal size, and coherent?
F. **Word selection principles**	Was there an explicit description of what would be counted and what would not be included? Were the criteria for inclusion (frequency, range, dispersion, or some composite measure) clearly described and justified?
G. **Manual checking**	Were any subjective criteria used? Were they described and justified?
H. **Ordering items**	Were the criteria for ordering in the list (frequency, range dispersion, or some composite measure) clearly described and justified? Were the criteria for making sub-lists clearly described and justified?
I. **Validation**	Were the lists checked against competing lists not just for coverage but also for overlapping and non-overlapping words?
J. Self-criticism	Are the weaknesses of the lists clearly acknowledged?
K. Availability	Are the lists readily available in electronic form for evaluation?

The ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018) were compared following eleven categories (A to K) in Table 3.2. Comparison tables were built to describe the main features of the two lists, followed by a detailed critique of each feature. Starting with the analysis of Nation's (2016) adapted framework helped to show how the lists were similar and different in terms of their features as well as the compilation process. This part, therefore, served as a foundation for further comparison of the lists.

Second, the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018) were compared in terms of lexical constituents using Venny (version 2.1.0) (Oliveros, 2015) to identify overlapping items between the lists. Items in the two lists with different spellings (e.g., *international organisation*

vs. *international organization*), with or without plural forms (e.g., *living condition* vs. *living conditions*), and with or without an optional article (e.g., *achieve (a) goal* vs. *achieve goal*) are treated as two different items by Venny. Therefore, manual checking was conducted to ensure that those items were counted as overlapping items.

Third, the two lists were compared in terms of coverage over independent corpora which were different from those that were used to develop the lists. The Academic and Fiction sub-corpora of the Corpus of Contemporary American English (COCA) developed by Davies (2012) were used in this study. The COCA Academic corpus is composed of 111 million words from different peer-reviewed journals in nine different disciplines. The COCA Fiction corpus contains 112 million words from magazines and movie scripts, which reflects non-academic language. These corpora were selected for three reasons. First, following Coxhead (2000), an academic corpus and a fiction corpus of comparable size can be used to evaluate the performance of academic word lists. Second, the COCA Academic and Fiction are currently the largest available corpora of similar sizes. Third, compared to other non-academic sub-corpora such as COCA Newspapers which contains factual texts on finance and politics, COCA Fiction features non-academic registers with personal and creative language.

The overall coverage, average coverage and coverage of the most frequent items over the COCA Academic and Fiction corpora were compared between the two lists. Because the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018) have a significant difference in the number of items and the overall coverage favours the longer list, the average coverage was also calculated for future reference. The coverage of the AECL (Lei & Liu, 2018) was further broken down into separate coverage for lexical items (hereafter, the Lex AECL) and grammatical items (henceforth, the Gram AECL) to have a better insight into the differences between the lists. The Gram AECL was excluded from the analysis of the most frequent items because the ACL contains lexical collocations only. The comparison could be clearer and more comprehensible when equal terms were compared. The ACL and the Lex AECL were then compared at three coverage points: the top 500, 1,000 and 2,469 academic collocations over COCA Academic. It is useful to look at the coverage changes at different cut-off points to see how the additional coverage changes as more items are added (Dang & Webb, 2016). The largest cut-off point of 2,469 items was selected based on the length of the shorter ACL

(Ackermann & Chen, 2013). The 1,000 and 500 cut-off points were comparison points for high-frequency multiword units.

To calculate the overall coverage of each list, the total frequency of all collocations in the list was first multiplied by two, assuming that collocations are made up of two words. This number was then divided by the total number of running words in the corpus (i.e., the corpus size), and multiplied by 100.

$$\text{Overall coverage} = \frac{\sum \text{frequency of all list items x 2}}{\text{Corpus size}} \text{ x } 100^{1}$$

To calculate the average coverage, the overall coverage of the lists was in turn divided by the number of items in each list:

$$\text{Average coverage} = \frac{\text{Overall coverage}}{\text{Number of list items}}$$

To calculate the coverage of the most frequent items of each list, the total frequency of those items was first multiplied by two, which was then divided by the corpus size and multiplied by 100:

$$\text{Coverage of most frequent items} = \frac{\sum \text{frequency of most frequent items x 2}}{\text{Corpus size}} \text{ x } 100$$

Antconc software (version 3.5.8.) (Anthony, 2019) was used to carry out the frequency analysis of the collocation. The advanced search function in the "Clusters/ N-Grams" tab in Antconc allows users to paste the entire list of multiword units. The results return the frequency of each item in the list and the total frequency of all items. This means that collocations were searched as fixed terms, and frequency data were based solely on one form of items in the original lists. It would be preferable to include all forms of one collocation (e.g., plural forms or past tense) in frequency counts. That said, this is an enormously time-consuming and difficult task, given the huge number of items in the

[1] I would like to thank an anonymous reviewer who suggested the following formula for calculating the coverage of multiword units (MWU): (MWU 1 x Number of words making up MWU 1 + ⋯ + MWU n x Number of words making up MWU n) / (Number of running words in the corpus) x 100. In the present study, the suggested formula has been adjusted with the assumption that all of the collocations in the two lists are made up of two words, despite the fact that some of them have optional articles.

two lists. Therefore, the data in this study were based only on one form of academic collocations in the published lists.

Results of the wordlist evaluation are presented in Chapter 4 (Section 4.2). The findings indicated that the ACL (Ackermann & Chen, 2013) was more suitable for the testing purpose in this project. Consequently, the item selection in Step 3 involves the ACL.

3.1.2 Domain modelling

The next three steps (Steps 3 to 5, Table 3.1) are based on the ECD's domain modelling layer. They involve sampling academic collocations from Ackermann and Chen (2013), choosing test format and writing test items. These steps were to create appropriate test tasks that enable test-takers to demonstrate their knowledge of academic collocations. It should be remembered that the Academic Collocation (AC) Tests in the present study consist of the AC Recognition Test and the AC Recall Test. These two tests focus on the same target academic collocations (Step 3) but employ different test formats (Steps 4 and 5) (see Chapter 4 for details).

3.1.3 Assessment implementation

The last step of piloting and finalising the ACTs (Step 6, Table 3.1) reflects the assessment implementation layer in the ECD. The AC Recognition Test and the AC Recall Test were piloted independently with participants who were similar to the target population of the ACTs. The main purposes of this stage were to pilot the test items and the scoring method. Based on the piloting results, the ACTs were finalised for the validation study.

The pilot AC Recognition Test and AC Recall Test were distributed online using Qualtrics as a testing platform. An invitation email was sent with the link of one of the tests to undergraduate and postgraduate students using English for academic purposes in Vietnam and New Zealand to ask for their voluntary participation. Students took the tests at their convenience and their results were automatically recorded by Qualtrics. A total of 79 students completed either of the tests (41 took the AC Recognition Test and 38 took the AC Recall Test). According to Wright and Tennant (1996), for a pilot study, results from 30 people are enough to run Rasch analysis (see Section 3.2.4.1). The data

collection for this piloting was approved by the Human Ethics Committee of Victoria University of Wellington, New Zealand (Reference number: 0000028145).

The results of the pilot ACTs were then examined using the Rasch model (Rasch, 1993) which was conducted with Winsteps Rasch software version 4.4.4 (Linacre, 2019). This model provides statistical tools to investigate score sets and highlight any problems with the ACTs. The analyses of fit indices were used to check the extent to which the data conformed to the Rasch model. If the data fit the model, then the most important use of the Rasch model is to select items in the pilot stage (Wu et al., 2016). The fit statistics are expressed in the form of mean-square values (MNSQ) as well as standardised values (ZSTD) with the generally accepted range for MNSQ is from 0.5 to 1.5 and for ZSTD is from -2 to +2 (Wright & Linacre, 1994). Items that have the MNSQ and ZSTD out of the above ranges are called misfit items. These items may be measuring an irrelevant construct and need to be re-investigated. Wright and Linacre (1994) provide guidance for interpreting MNSQ values (see Table 3.3). The fit statistics help to detect problematic items, but only when sources of misfit can be identified, then the decision should be made whether items need removing or improving (Wu et al, 2016). Based on this analysis, the best test items were selected for the final test version used in the validation process.

Table 3.3

Interpretation of Mean-Square Fit Statistic Values (Wright & Linacre, 1994)

MNSQ value	Interpretation
> 2.0	Distorts or degrades the measurement system
1.5 – 2.0	Unproductive for construction of measurement, but not degrading
0.5 – 1.5	Productive for measurement
< 0.5	Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations.

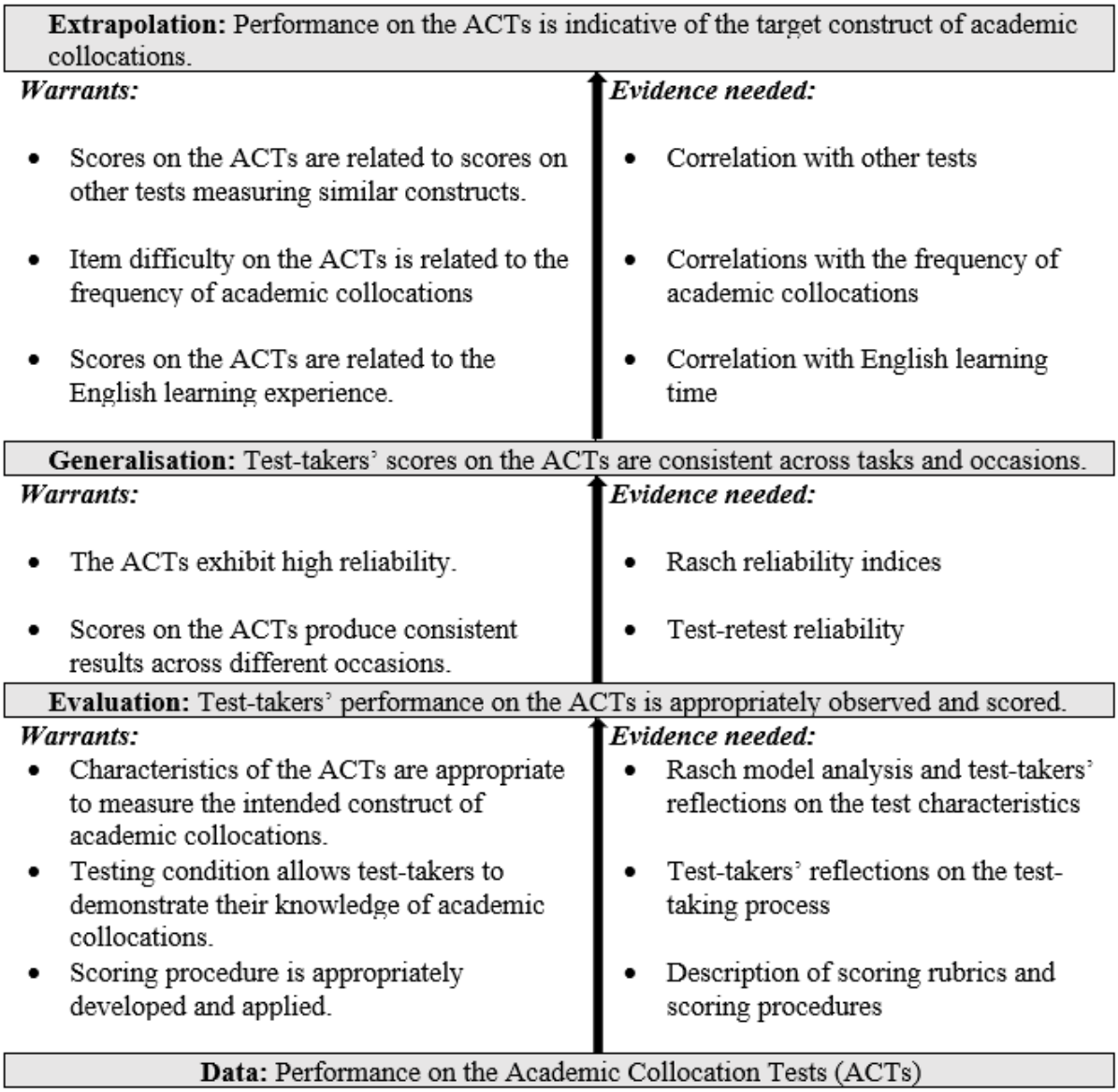
3.2 Validation process for the ACTs

The present study follows the argument-based approach (Kane, 1992, 2004, 2013) to validation of the ACTs. This study adopted an exploratory mixed-methods design (Creswell & Clark, 2017) in which both quantitative and qualitative data were gathered for the assessment of the inferences in the validation framework of the ACTs. Figure 3.1 summarises the inferences and the evidence needed

that guided the data collection. As explained in Chapter 2 (Section 2.5.2), only three inferences (Evaluation, Generalisation and Extrapolation) are included in the validation framework for the ACTs.

Figure 3.1

Argument-Based Validation Framework for the ACTs



It is important to note that the test development and validation are complementary, with evidence for the inferences in the validation framework of the ACTs being collected both during and after the test

development. For example, the description of the scoring system as a backing for the Evaluation inference is presented in Chapter 4 (Section 4.6.1) as part of the test development process. Other evidence was collected through an empirical study whose participants, materials and procedures are detailed below.

3.2.1 Participants

A convenience sampling method was employed so that data could be collected from a large number of participants in Vietnam and New Zealand. To be included in this study, participants had to have experience of formal English study at a university either in Vietnam or in New Zealand. University students in Vietnam represented the English as a Foreign Language (EFL) population, while students in New Zealand were representatives of the English as a Second Language (ESL) population. The data collection for this validation study was approved by the Human Ethics Committee of Victoria University of Wellington, New Zealand (Reference number: 0000028145).

In Vietnam, a total of 233 Vietnamese students completed the tests and answered a background questionnaire (see Section 3.2.2.1), including 211 females and 22 males in two different universities. Out of these participants, 24 students (i.e., three males and 21 females; 10% of the total number of participants in Vietnam), took part in a post-test interview to share their opinions about the ACTs and reflections about the test-taking process (see Section 3.2.2.3). The two universities were selected because they were among the first who could switch to online learning in response to COVID-19 pandemic. It was an unexpected event that affected the teaching and learning routines of universities in Vietnam. It took time to set up and train students and staff before the universities could entirely move to online learning. The students' age ranged from 18 to 23, with an average age of 20. Their majors included English linguistics, English translation and interpretation, English language teacher education and primary English language teacher education. This means they were all English-major students who on average had already had more than nine years of formal English learning in Vietnam. At the time of data collection, the participants were in the second year (N = 128) and third year (N = 105) of university studies. First-year and fourth-year students were not recruited because during this time an e-learning system had not been completely developed for these two groups and it was therefore difficult to contact these students. First-year students were in an exam-preparation

period, and they had not started a new trimester yet, while fourth-year students were taking internships.

In New Zealand, participants were all non-native English speakers who had been studying in an English-speaking country for at least a year. A total of 110 students completed the online survey and testing, including 87 females, 22 males and one non-binary participant. A total of 51 were Vietnamese and the other 59 participants came from a wide range of first language backgrounds such as Chinese, Filipino, Thai, German, Spanish, Malay and Hindi. The majors of this ESL group also ranged widely: Linguistics and Applied Linguistics (18), TESOL (11), Education (8), Economics and Finance (4), Psychology (4) and others (65). There were a range of students in this study including undergraduate (30), postgraduate (55) and others (25). A total of 20 out of 110 students in this group participated in the post-test interview (18 females and 2 males).

3.2.2 Materials and instruments

This study employed four data collection instruments, namely, a background questionnaire, the ACTs, the Vocabulary Size Test (Nation & Beglar, 2007) and the post-test interview. The development of the ACTs will be reported in the next chapter. The other three instruments are now described in turn.

3.2.2.1 Background questionnaire

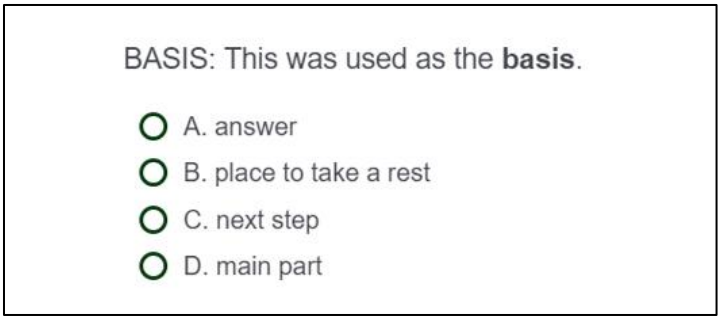
A questionnaire was developed to obtain the participants' background information including name, gender, age, first language, study level and major, length of learning English and length of studying in an English-speaking country. The questionnaire version for the participants in New Zealand had two additional screening questions and two questions related to English language proficiency test results (i.e., IELTS or TOEFL scores). In Vietnam, these questions were skipped because they were not applicable to Vietnamese students. The purpose of the questionnaire and brief instructions were presented before the questions. The questionnaire was written in English and was piloted by two ESL students in New Zealand before being made available online using Qualtrics software. The full version of the questionnaire is in Appendix A.

3.2.2.2 The Vocabulary Size Test (VST)

The Vocabulary Size Test (Nation & Beglar, 2007) is a measure of a learner's overall English vocabulary size (i.e., how many English words that he or she recognises). It has a total of 140 items in a multiple-choice format. Test-takers were presented with a target word in a non-defining context and four different definitions of the word to select from. An example of a VST test item is displayed in Figure 3.2. Each test item represents 100 word families. In other words, every 10 items in the test represent one 1,000-word level out of 14 levels based on a frequency count of word families in the British National Corpus. The VST score is obtained by multiplying the test score by 100 to give an estimate of a learner's vocabulary size. For example, a learner with a score of 75 out of 140 on the test has a vocabulary size of 7,500 word families. The VST has been found to be a reliable and valid instrument to measure vocabulary size according to a validation study using Rasch analysis based on Japanese test takers (Beglar, 2010).

Figure 3.2

Example of a Vocabulary Size Test Item (Nation & Beglar, 2007)



Note. The test is publicly available on Paul Nation's website at

https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests

The VST was selected to be used in this study instead of the Vocabulary Levels Test (VLT) (Schmitt et al., 2001) because the VST has a wider coverage of frequency bands (from 1,000 to 14,000 word levels), while the VLT only includes several sample levels (2,000; 3,000; 5,000 and 10,000 word levels). Even in the later version, the updated VLT (Webb et al., 2017) adds test items for levels 1,000 and 4,000, but not for other levels. Due to this feature, Nation and Beglar (2007) argue that the

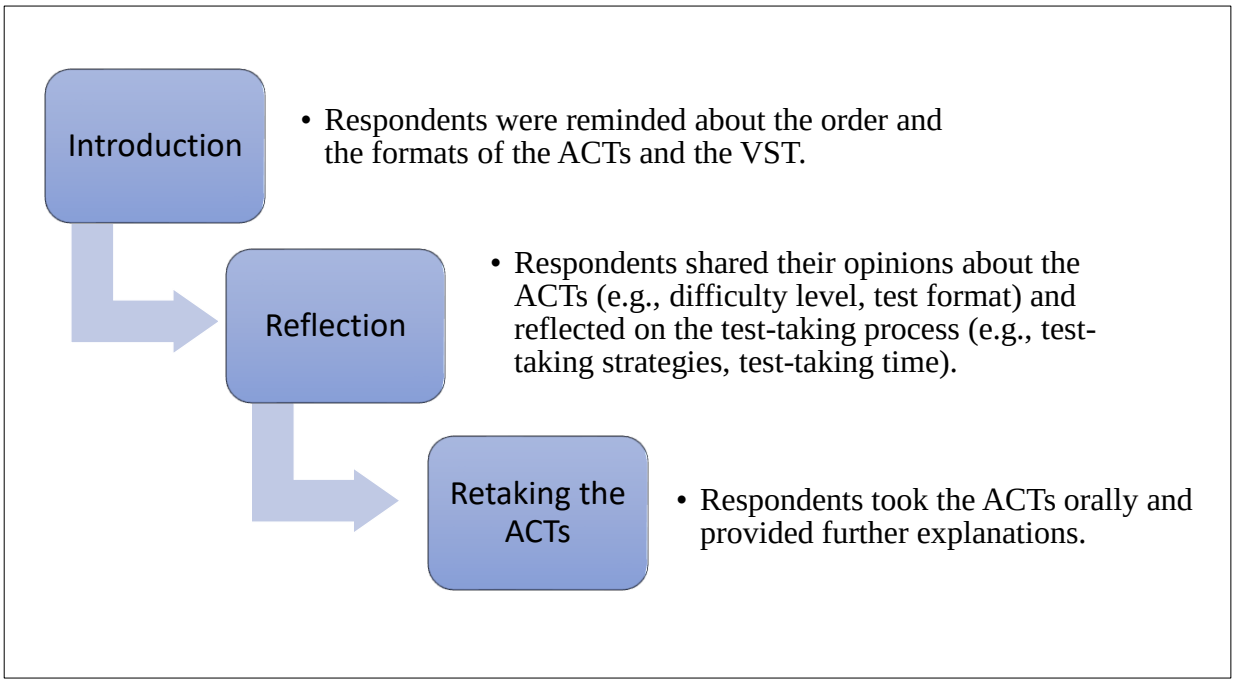
VST is closer to a vocabulary proficiency test and the VLT is a diagnostic measure (see more in Stoeckel, McLean & Nation, 2021 for a systematic distinction between the VST and the VLT).

3.2.2.3 Interview questions

The aims of the interview were to have the test-takers reflect on the ACTs and to explore whether the answers given on the online tests would correspond with their verbalised knowledge. This method, known as retrospective verbal report (Cohen, 1998), was also used to validate the VLT by Schmitt et al. (2001). The interview in the present study included three parts: introduction, reflection and re-taking the ACTs verbally. The purpose of each part is summarised in Figure 3.3. The outline of the interview with specific questions is in Appendix B.

Figure 3.3

Summary of Post-Test Interview Sections



3.2.3 Data collection

In the original pre-COVID plan, the tests and the questionnaire were to be administered in Vietnam and then in New Zealand under my supervision in a computer lab on university campuses. However,

universities in Vietnam were closed between February and March 2020, while universities in New Zealand also went online in March 2020. As a result, the method of data gathering was changed from paper and pen to online, using Qualtrics as a survey and testing platform. In Vietnam, information about the research with the online survey and testing link was delivered to students by their English teachers (who had been contacted by email). In New Zealand, participants were recruited via social media sites (i.e., online Facebook groups of students). Participants contacted the researcher via email to receive the link to the survey and tests. The informed consent form was attached to the online survey and testing link stating that students' participation was voluntary. Students took the survey and tests at their convenience and there was no time limit. It was unlikely that participants would start the test battery from the beginning if they could not finish all the tests in one attempt. A time limit, therefore, was not set for test-takers to increase the chance of the tests being completed. All the participants took the online survey and tests in the following order, noting that a reminder to take a five-minute break between the tests automatically appeared at the end of the first and the second test.

(1) Questionnaire
(2) The AC Recall Test
(3) The Vocabulary Size Test
(4) The AC Recognition Test

Once the participants had completed the online survey and testing, an invitation email with the information sheet for the post-test interview was sent to them. The consent form was then sent to those who agreed to participate. The semi-structured interviews were conducted within two weeks after the participants had taken the online tests. The interviews took place in a neutral, unthreatening environment via Zoom – an online video and audio communication platform. All the interviews were conducted one-on-one in Vietnamese or English and were audio-recorded with the permission of the interviewees. The length of the interviews varied from 30 minutes to one hour.

3.2.4 Data analysis

This section outlines how different types of analysis were conducted with different data sets (Table 3.4). It is worth noting that for the main data analysis, all the data were pooled together instead of separating out the Vietnamese and New Zealand data. This is because a large group of participants

was needed to run the statistical model in order to reliably investigate the performance of the ACTs. The test results for participants in two different contexts are still reported later in Chapter 6 (Section 6.2.1) for future reference. The data analysis presented in this section aims to answer the research questions and provide evidence pertaining to the inferences in the validation framework of the ACTs. Table 3.4 summarises the alignment between the validity inferences and the research questions, as well as the methods for data analysis.

Table 3.4

Summary of Data Analysis in Relation to Research Questions and Validity Inferences of the ACTs

Inference	Research question	Type of evidence	Analysis method
Evaluation	RQ1: To what extent the characteristics of the ACTs are appropriate to measure the intended construct of academic collocations?	Test scores	Rasch analysis (Section 3.2.4.1)
		Interview data	Thematic analysis (Section 3.2.4.3)
	RQ2: Does the testing condition allow test-takers to demonstrate their knowledge of academic collocations?	Interview data	Thematic analysis (Section 3.2.4.3)
Generalisation	RQ3: Are scores on the ACTs reliable?	Test scores	Rasch analysis (Section 3.2.4.1)
			Correlation analysis (Section 3.2.4.2)
Extrapolation	RQ4: Are scores on the ACTs related to scores on other tests measuring similar constructs?	Test scores	Correlation analysis (Section 3.2.4.2)
	RQ5: Is the item difficulty on the ACTs related to the frequency of academic collocations?	Test scores and corpus frequency	
	RQ6: Are scores on the ACTs related to English learning experience?	Test scores and survey data	

3.2.4.1 Rasch analysis

Initially, the results of the AC Recognition Test, the AC Recall Test and the VST were scored, and the data were imported into Excel spreadsheets and IBM SPSS (version 25) for analysis. The AC Recognition Test and the VST were scored automatically by Qualtrics, and the AC Recall Test was scored by me (see Chapter 4, Section 4.6.1 for details about the scoring system). After that, basic

information on each test, including a minimum score, a maximum score, an average score and a standard deviation was reported. Descriptive statistics helped to visualise test-takers' performances on the tests.

The results of the AC Recognition Test and the AC Recall Test were exported to Winsteps software version 4.4.4 (Linacre, 2019) for Rasch model analysis. This is a statistical method for analysing the score sets and determining whether the test items performed well to measure the target construct. Following this model, five properties of the tests were examined:

- **Item measure**: Under the Rasch model, logit used as the measurement unit shows the difficulty level of test items (higher logit value means harder item). The item difficulties of test items are examined in relation to person abilities (also measured in logits). Wright maps are built to visualise the relationship between test-takers' knowledge and item measures on the same scale of measurement.
- **Item fit**: If test items fit the Rasch model, it can be concluded that they involve one single construct and each item fits with the other items to measure that only one construct (Bond & Fox, 2007). The fit statistics mean squares (MNSQ) and the standardised Z (ZSTD) are used to assess whether an item functions as the Rasch model expects. The expected MNSQ value of the fit statistics is 1.0. The MNSQ value between 0.5 and 1.5 is generally regarded as an acceptable fit to the Rasch model (Linacre, 2002). For mid- or low-stakes multiple-choice questions, a more stringent range of 0.7 to 1.3 is suggested (Wright & Linacre, 1994). Meanwhile, the ZSTD values ranging from -2.0 to 2.0 are suggested to indicate an acceptable fit to the Rasch model (Linacre, 2002). According to Bond and Fox (2007), the four types of fit statistics (infit MNSQ, outfit MNSQ, infit ZSTD, and outfit ZSTD) can be used independently or in combination to judge the item fit, but more emphasis should be put on infit rather than outfit values. This is because infit statistics focuses more on test-takers' performances close to the item difficulty, while outfit statistics is more sensitive to outliers such as random guessing or careless responses. Infit statistics were thus used in the present research to detect misfit items.

  There are two types of misfit items: underfit and overfit. Underfit refers to items that contain more unexpected responses than the model could predict and therefore may degrade the

quality of the measurement. Underfit is identified when MNSQ is greater than a particular value (e.g., 1.5 or 1.3) or ZSTD is greater than 2.0. Overfit, on the other hand, means the items show less variability than predicted, so the reliability may be overestimated (Wright & Linacre, 1994). Overfit is indicated as MNSQ less than a particular value (e.g., 0.5 or 0.7) or ZSTD less than -2.0. The underfit is hence of greater concern.

- **Point-measure correlations**: Rasch model calculates correlations between item responses and test-takers' knowledge. Positive correlations are expected so that test item responses correspond to people's knowledge. Items with a negative or low correlation (r <.20) may require additional investigation (Linacre, 2019).
- **Unidimensionality**: This is an important assumption of the Rasch model, which presumes that a test measures one single trait or construct (Fan & Bond, 2019). Principal component analysis of residuals (PCAR) is conducted to address the dimensionality issue. This analysis helps to detect whether any other unintended constructs is being measured (i.e., secondary dimensions). In PCAR, the first contrast is checked whether it is substantive enough to represent a component. A first contrast with an eigenvalue above 2 suggests a possible presence of an additional dimension (Linacre, 2019).
- **Local independence**: In Rasch measurement, the items are required to be independent of each other. That is, a correct/ wrong answer to one item should not lead to a correct/ wrong answer to another item. This local independence is identified by Rasch residual correlations via Q3 coefficients. Local independence can be confirmed with small values of Q3 coefficients between -0.3 and 0.3 (Fan & Bond, 2019). The local independence and the unidimensionality are two interrelated properties. Linacre (2019) provides the following rule of thumb: "All items must be about the same thing, our intended latent variable, but then be as different as possible, so that they tell us different things about the latent variable. But when two or more items tell us the same "different thing", then we have indications of a secondary dimension" (p.19).

These five properties of the tests are the most important metrics of validity suggested by researchers in language testing (e.g., Aryadoust et al., 2020; Fan & Bond, 2019). Results of the ACT item

analysis are presented in Chapter 5 and are used to assess the Evaluation inference in the validation framework of the ACTs (Figure 3.1).

The AC Recognition Test and the AC Recall Test were then investigated in terms of reliability. Winsteps software provided different Rasch reliability metrics, consisting of reliability and separation indices for items and persons. These indices helped to determine the extent to which the test results could be considered consistent and how much the tests could distinguish learners at different levels. The reliability results are presented in Chapter 6 as evidence for the assessment of the Generalisation inference in the validation framework of the ACTs (Figure 3.1).

3.2.4.2 Correlation analysis

As the score sets were not normally distributed (see descriptive statistics in Chapter 5), Spearman's correlation (r_s) – a non-parametric measure of rank correlation – was employed to indicate the relationship between different data sets in the present study. First, correlations between the AC Recognition Test and the AC Recall Test, as well as between these two tests and the VST, were calculated. Second, Spearman's correlation was used to determine whether there was a relationship between scores on the ACTs and the frequency of academic collocations. Third, Spearman's correlation was employed to correlate the ACT scores with survey data, including IELTS scores and years of English learning. Results of these correlation analyses are presented in Chapter 6 and are used for the assessment of the Extrapolation inference in the validation framework of the ACTs (Figure 3.1).

The post-test interview also provided quantitative data when the participants re-took the tests verbally. In the present study, the test-retest reliability was indicated by Spearman's correlation coefficient (r_s) which measures the strength of the relationship between the two data sets. If the tests could consistently produce the same results, then the relationship between the two score sets (pre and post) should be high. The test-retest reliability was presented in Chapter 6 to assess the Generalisation inference in the validation framework of the ACTs (Figure 3.1).

3.2.4.3 Thematic analysis

The next data analysis step addresses the qualitative data from the interviews which were analysed to investigate test-takers' opinions about the ACTs and their reflections on the test-taking process. Thematic analysis (Braun & Clarke, 2006, 2012) was applied to examine the interview data with six steps: 1) Familiarising myself with my data, 2) generating initial codes, 3) searching for themes, 4) reviewing themes, 5) defining and naming themes, and 6) producing the report.

Taking the first step of Braun and Clark's (2006, 2012) approach, I familiarised myself with all audio-recorded interviews through organising, transcribing, and translating the data. A total of 44 recordings with an overall duration of around 29 hours were put in one folder and stored safely. I transcribed the first part of the interviews where the participants were asked to make reflections and give opinions about the ACTs. For the second part of retaking the tests verbally, I took notes on the respondents' answers for all the test items, which were used for calculating the test-retest reliability (see Section 3.2.4.2). I also transcribed further comments or explanations (if any) from the participants. I then translated the interviews in Vietnamese into English. For cross-checking, I invited a bilingual Vietnamese PhD candidate in New Zealand to check my translations of all the texts quoted in this thesis. This cross-checking was approved by the Human Ethics Committee of Victoria University of Wellington, New Zealand (Reference number: 0000028145).

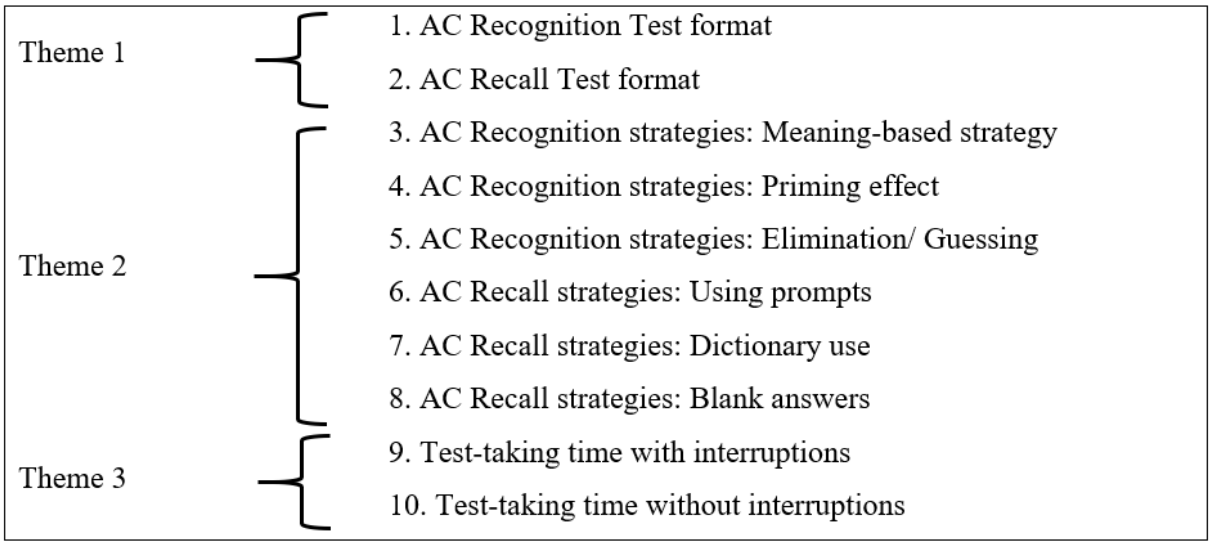
Following a thorough reading of the interview scripts, I generated initial codes and themes as the second and third steps. The validation framework of the ACTs (Figure 3.1) was employed to search for themes within the data. For example, the Evaluation inference needs evidence to support the test format; therefore, segments of the raw transcribed text with keywords such as "format" or "test" were highlighted and collated under one theme.

In the next two steps, I reviewed, defined and named the themes. The three emerging themes from the analysis are 1) test formats, 2) test-taking strategies and 3) test-taking time. These themes were divided into ten coded categories as presented in Figure 3.4. To ensure the robustness of the coding system, a Vietnamese researcher in applied linguistics who is knowledgeable in language testing was invited to act as an independent coder. This coder was trained before she started working on 10% of the data set. Her work was then compared to mine, and the inter-coder reliability was calculated,

which yielded the result of 91% of agreement. The independent coder and I then further discussed the differences between our coding until there were no disagreements.

Figure 3.4

Categories for Coding Interview Data



Finally, the findings from the interview data were used to assess the Evaluation inference in the validation framework of the ACTs (Figure 3.1). These findings are reported in Chapter 5. To ensure confidentiality and anonymity, the respondents are identified by pseudonyms with a tag of "VN" or "NZ" (e.g., Trang NZ) to clarify whether the participants were from Vietnam (VN) or New Zealand (NZ) context. The quotations selected to be presented in this thesis are either typical responses that represented opinions from others or important aspects of the ACTs acknowledged by the respondents. The square brackets […] are used to clarify detail where necessary.

3.3 Chapter summary

This chapter presented the methodology used for developing and validating the ACTs. For the test development process, the six steps of creating the ACTs based on a corpus-based word list were outlined. Among them, the methods for evaluating the academic collocation lists to select the best item source and the piloting stage of the tests were highlighted. For the validation process,

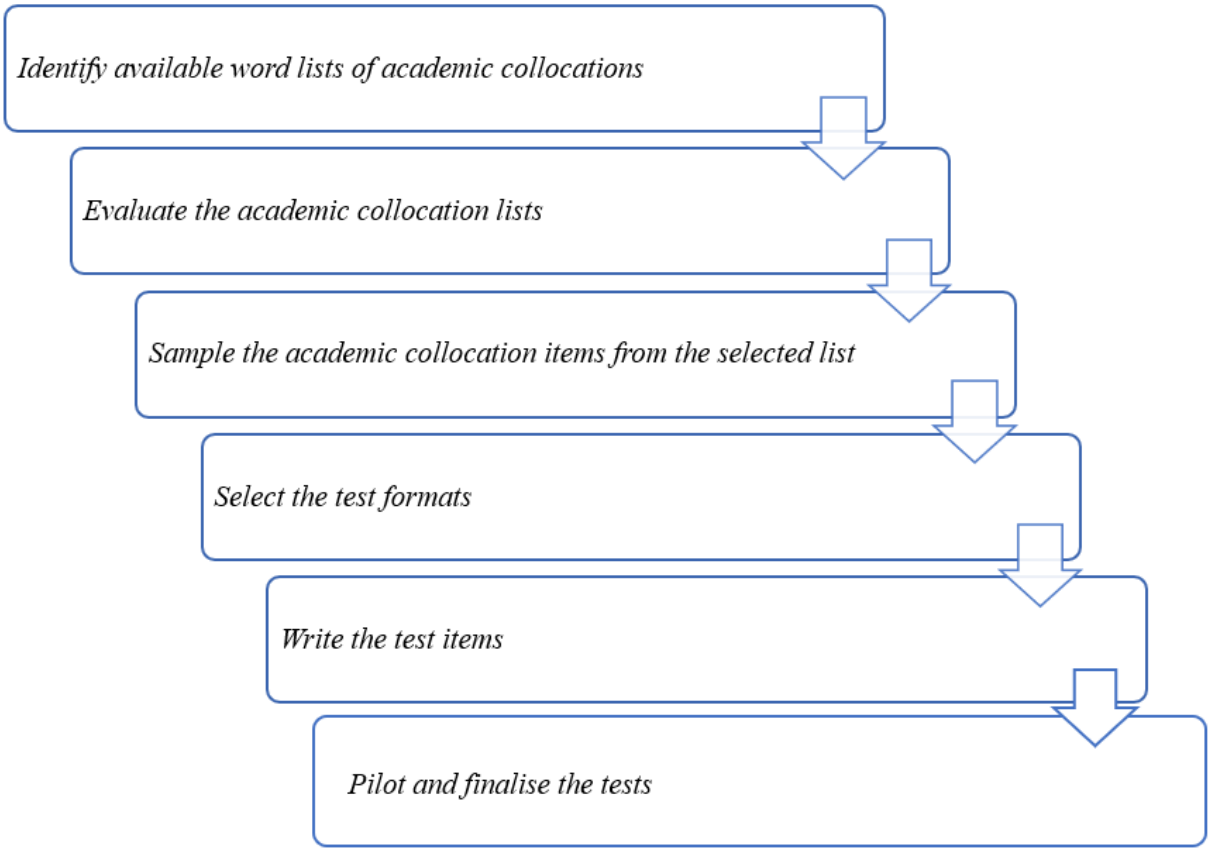
participants, materials and procedures employed in the empirical research to collect validity evidence for the ACTs were described. The next chapter reports in detail the development of the ACTs.

Chapter 4 Development of the Academic Collocation Tests

The Academic Collocation Tests (ACTs) consist of two tests: the AC Recognition Test to measure learners' ability to recognise academic collocations in a given context and the AC Recall Test to measure learners' ability to produce academic collocations in a given context. The process of developing these tests from a corpus-based word list involves six steps (see Figure 4.1). These steps are based on the evidence-centred design framework (Mislevy & Yin, 2013), as discussed in Chapters 2 and 3. Sections 4.1 to 4.6 in this chapter each focus on a different step in the test development process, with Section 4.7 serving as a chapter summary.

Figure 4.1

Steps of Developing the ACTs From a Corpus-Based Word List



4.1 Identifying available word lists of academic collocations

The first step in creating the ACTs was to identify potential sources of academic collocations from which the test items could be sampled. Instead of creating a new list, I decided to make use of a pre-existing one. This was because a well-made word list requires a great deal of time and effort, and word lists are seen as an avenue for testing (Nation, 2016). As there were several published lists of academic collocations, a better option could be making use of one of these sources for developing the ACTs. Chapter 2 (Section 2.3.1) identified four lists of academic collocations: Durrant (2009), Chon and Shin (2013), Ackermann and Chen (2013) and Lei and Liu (2018). Durrant (2009) and Chon and Shin (2013) contain mostly grammatical collocations which are incomplete expressions such as *and respectively* or *of income*. These collocations are less salient and seem to have less pedagogical value than those with complete meanings in the lists of Ackermann and Chen (2013) and Lei and Liu (2018) (e.g., *significant difference, relatively high*). Therefore, the Academic Collocation List by Ackermann and Chen (2013) and the Academic English Collocation List by Lei and Liu (2018) were selected for further investigation and evaluation.

4.2 Evaluating the Academic Collocation List (ACL) (Ackermann & Chen, 2013) and the Academic English Collocation List (AECL) (Lei & Liu, 2018)

The second step for the test development was to evaluate the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018). This step is important for selecting the better candidate between the two lists for the creation of the ACTs. As discussed in Chapter 3 (Section 3.1.1), these lists were evaluated by using an adapted framework from Nation (2016), comparing lexical constituents and analysing lexical coverage. The evaluation results are in turn presented below.

4.2.1 Applying Nation's (2016) adapted framework to the ACL and the AECL

Nation's (2016) adapted framework (see Section 3.1.1) was broken down into smaller sections for a comprehensive analysis. The ACL and the AECL were compared in terms of the target audience, purpose, number of items, unit of counting, corpus, word selection principles, manual checking, ordering items, validation, self-criticism and availability. The main features of the two lists are presented in tables with a detailed commentary on each.

Table 4.1

Audience and Purpose of the ACL and the AECL

Word list	The ACL (Ackermann & Chen, 2013)	The AECL (Lei & Liu, 2018)
A. Target audience	EAP teachers and learners	ESL/EFL students and professionals learning and/ or using academic English
B. Purpose	To develop a list of the most frequent and pedagogically relevant collocations in academic English that is useful for EAP teaching, learning and assessment.	To create a type-balanced list of the most frequent English academic collocations that can be used as a useful reference and teaching resource.

A. Target audience

As can be seen from Table 4,1, the target audiences of the two lists are clearly described. Both lists aim at EAP teachers and learners. However, the target audience of the AECL (Lei & Liu, 2018) may be a bit wider to include practitioners and researchers who may use the list for research purposes.

B. Purpose

In terms of purposes, Ackermann and Chen (2013) stress the pedagogical value of their list, while Lei and Liu (2018) add an additional focus on the type-balanced feature of the list, which means a wide range of collocation kinds are included in the list with a reasonable ratio division.

Table 4.2

Number of Items, Unit of Counting and Corpora of the ACL and AECL

Word list	The ACL (Ackermann & Chen, 2013)	The AECL (Lei & Liu, 2018)
C. Number of items	2,469	9,049
D. Unit of counting	No information provided	Lemmas
E. Corpora	The written component of the Pearson International Corpus of Academic English (PICAE)	A corpus created from six different corpora: the British National Corpus, the British Academic Written English Corpus, and the Jiao Da English for Science and Technology Corpus; and three corpora compiled by the authors: journal articles, doctoral dissertations and book reviews
• Size	25.6 million words	43.1 million words
• Corpus texts	333 documents including journal articles and textbooks	7,214 texts including academic book snippets, articles in periodicals published/ unpublished manuscripts, academic writing by graduate and undergraduate students, theses from universities in English-speaking countries, textbooks, and articles and book reviews from international journals
• Academic disciplines	28 academic subjects which are divided into four academic disciplines (seven subjects per discipline): applied sciences and professions, humanities, social sciences, natural/formal sciences	Five discipline divisions: natural sciences, applied sciences/engineering, social sciences, applied social science, and humanities

C. Number of items

Table 4.2 shows that the AECL (Lei & Liu, 2018) has nearly four times as many items as the ACL (Ackermann & Chen, 2013), which means that for pedagogical purposes, the AECL has several disadvantages compared with the ACL. For teaching purposes, the lists have to be made as practical and usable as possible for EAP teachers and learners. One list of more than 9,000 items is huge and clearly over the limit of practicality. A lengthy list poses a greater challenge in making decisions

about which items to focus on to maximise the benefit within the limited classroom time (Simpson-Vlach & Ellis, 2010). Earlier lists for pedagogical purposes also limit their number of items to a manageable data set, for example, the Academic Word List (Coxhead, 2000) (570 word families) or the Academic Formulas List (Simpson-Vlach & Ellis, 2010) (200 formulas for written language, 200 for spoken language, and a combined list of 207 formulas for written and spoken language). That said, the length of the AECL can be explained when another function of the list is considered. Lei and Liu (2018) intended the list also to be a reference resource for practitioners, researchers and ESL learners.

D. Unit of counting

The unit of counting is one of the most important considerations when making the list of single words because different ways of counting a word may result in different word lists (Gardner, 2007). This feature does not seem to be as straightforward in lists of multiword units compared to single words. It appears that Ackermann and Chen (2013) used lemmas as the unit of counting because they state that they removed inflections and listed collocations in their base form (e.g., *professional activities* → *professional activity*). Lei and Liu (2018) chose to use lemmas in the AECL as the unit of counting because they believe that the pattern of a collocation does not change when inflectional suffixes are added to the stem (e.g., *make/makes/made/making a decision/ decisions*).

The analysis in this thesis showed that both lists show inconsistencies because plural forms (e.g., *living conditions*) and comparative forms (e.g., *broader context*) were found in the lists. These examples suggest that it is likely that types rather than lemmas could be the unit of counting of the two lists. According to Shin and Nation (2008), types should be used as the counting unit for collocations rather than lemmas or families because different types of the same word family have different collocates. For example, *widely* frequently collocates with the verb participle *distributed*, but not the other types of the verb *distribute*. Ackermann and Chen (2013) and Lei and Liu (2018) could perhaps have clarified the operational definition of what would be counted as an academic collocation and consistently applied that definition, as suggested by Nation et al. (2016).

E. Corpora

It is important that the corpus selection should be derived from the purpose of the word list (Nation & Webb, 2011). As already mentioned above, both the ACL and the AECL target EAP learners and teachers and aim to provide a pedagogically relevant list; therefore, any corpora should contain texts for academic purposes that represent the kinds of reading that university students are expected to do, across a range of subjects. To look closer into features of a corpus, Nation and Sorell (2016) suggest examining corpus size, text types and proportions of disciplines. Let us look at each of these points in turn.

Table 4.2 shows that the corpus from the AECL (Lei & Liu, 2018) is almost twice the size of the ACL's (Ackermann & Chen, 2013) corpus (25 million to 43 million). With respect to the question of how large a corpus should be, there are different opinions in the literature. A corpus size of around 20 million words is suggested by Brysbaert and New (2009) and Sorell (2013) to obtain a reliable word list of low-frequency words. Brysbaert and New (2009) posit that a corpus of between 16 and 30 million words can provide reliable word frequency norms for most practical purposes. Following suggestions by Brysbaert and New (2009) and Sorell (2013), both the ACL and the AECL are based on corpora that are large enough to provide sufficient instances and reliable results. That said, caution should be used when interpreting the reliability of corpus size as there are still no accepted guidelines.

Both studies use a variety of published academic manuscripts such as journal articles and textbooks, but with one difference in text types. The AECL (Lei & Liu, 2018) draws on the British Academic Written English (BAWE) corpus (Nesi & Gardner, 2012), which is a collection of proficient university-level students' writing. The ACL (Ackermann & Chen, 2013) corpus does not include student writing. Previous studies (e.g., Altenberg & Granger, 2001; Li & Schmitt, 2010) point out that even advanced L2 learners have problems with collocations used in academic writing. Published research articles, on the other hand, provide "a suitable basis for identifying pedagogically-useful academic collocations […] because of its centrality within academic writing (Durrant, 2009, p.159). Hyland (2008) further notes that research articles are considered "a model of good academic writing and as an ideal to be emulated as far as possible" (p.47).

Proportions of academic disciplines in the ACL and the AECL corpora are displayed in Table 4.3. Both corpora include a variety of disciplines of different proportions.

Table 4.3

Discipline Divisions in the Corpora of the ACL and the AECL

Discipline	ACL	AECL
Applied science	31%	22.23%
Applied social science	{25%}	16.64%
Social science		24.45%
Natural science	23%	15.56%
Humanities	21%	21.11%

The corpus of the ACL (Ackermann & Chen, 2013) contains a higher combined amount of natural and applied science (54%) than the other disciplines (46%). On the other hand, Lei and Liu's (2018) corpus contains a higher proportion of social sciences and humanities (62.20% compared to 37.79%). Proportions of the disciplines are relatively equal in the ACL but skewed to humanities and social sciences in the AECL. If the lists are intended to be equally useful for students irrespective of their field of study, it is important to keep equal proportions among the disciplines as much as possible (Coxhead, 2000; Nation, 2016). This is not an easy task.

F. Word selection principles

Both Lei and Liu (2018) and Ackermann and Chen (2013) began creating their lists by choosing content words as node words for extraction from their corpora. The AECL (Lei & Liu, 2018) was based on the Academic Vocabulary List (Gardner & Davies, 2014) as node words. In Ackermann and Chen's (2013) study, content words excluding high-frequency words from the General Service List (GSL) (West, 1953) was used as node words. Justifications for node word selection in the ACL are not provided. It is likely that Ackermann and Chen (2013) follow Coxhead (2000) to exclude words in the GSL with the assumption that EAP students already have knowledge of these general high-frequency words. More recent research (Gardner & Davies, 2014; Nation, 2016) has pointed out that academic vocabulary can occur be high, mid or low frequency vocabulary, so the case could be made also for academic collocations to include high, mid and low frequency vocabulary.

Table 4.4

Selection Principles of the ACL and the AECL

Word list	The ACL (Ackermann & Chen, 2013)	The AECL (Lei & Liu, 2018)
F. Word selection principles		
• Node words	Content words occurring at least five times per million words and in at least five different texts, excluding words from the GSL (West, 1953)	3,015 core words in the Academic Vocabulary List (Gardner & Davies, 2014) were used as node words
• Kinds of collocation	8 kinds: • Adjective + noun • Noun + noun • Verb + noun • Verb + adjective • Adverb + verb • Verb + adverb • Adverb + verb participle • Adverb + adjective	11 kinds: • Adjective + noun • Verb + noun • Adverb + verb • Noun + noun • Noun + verb • Adverb + adjective • Noun + preposition • Verb + preposition • Adjective + preposition • Verb + adjective • Adverb + adverb
• Criteria for selection of collocations	5 criteria: • Occurring at least five times in total across at least five different texts • Normed frequency ≥ 1 per million • Normed frequency ≥ 0.2 per million in each field of study • MI score ≥ 3 • T score ≥ 4	4 criteria: • Minimum frequency: 10 occurrences per million words • MI score ≥ 3 • T score ≥ 2 • Occurring 0.2 per million words in each of the five discipline divisions
G. Manual checking	• Manual vetting by the two authors • Expert review	The authors refine the list without expert review.
H. Ordering items	Collocations are ordered alphabetically.	Nodes are alphabetically ordered with all possible collocates. Therefore, many items are listed twice.

The next point is the inclusion of kinds of collocations. As can be seen from Table 4.4, collocations in the ACL (Ackermann & Chen, 2013) are divided into eight categories while the AECL (Lei & Liu, 2018) has eleven categories. The AECL includes prepositions-based patterns of collocations, which means this list contains both lexical and grammatical collocations. The AECL claims to be a more balanced list in its kinds of collocations, but two points need to be noted here. The first is that Lei and Liu (2018) do not discuss the distinction between lexical and grammatical collocations, and why only prepositions-based patterns of collocations are included but not the other kinds of grammatical collocations. Secondly, even though Lei and Liu (2018) stress the importance of prepositions-based collocations, verb + preposition and adjective + preposition collocations do not appear in the *LTP Dictionary of Selected Collocations* (Hill & Lewis, 1997) which was used to validate the list.

The ACL seems to be more straightforward with its explicit focuses on lexical collocations only. It also excludes some collocation kinds as in the AECL because of the seeming lack of pedagogical value. For example, noun + verb collocations are not included in the ACL but many are listed in the AECL, which may confuse learners because these items look grammatically incorrect (e.g., *colleague show* or *combination produce*). In sum, in having a fewer number of collocation kinds, the ACL (Ackermann & Chen, 2013) appears to be more selective in choosing kinds of collocations for their pedagogical list.

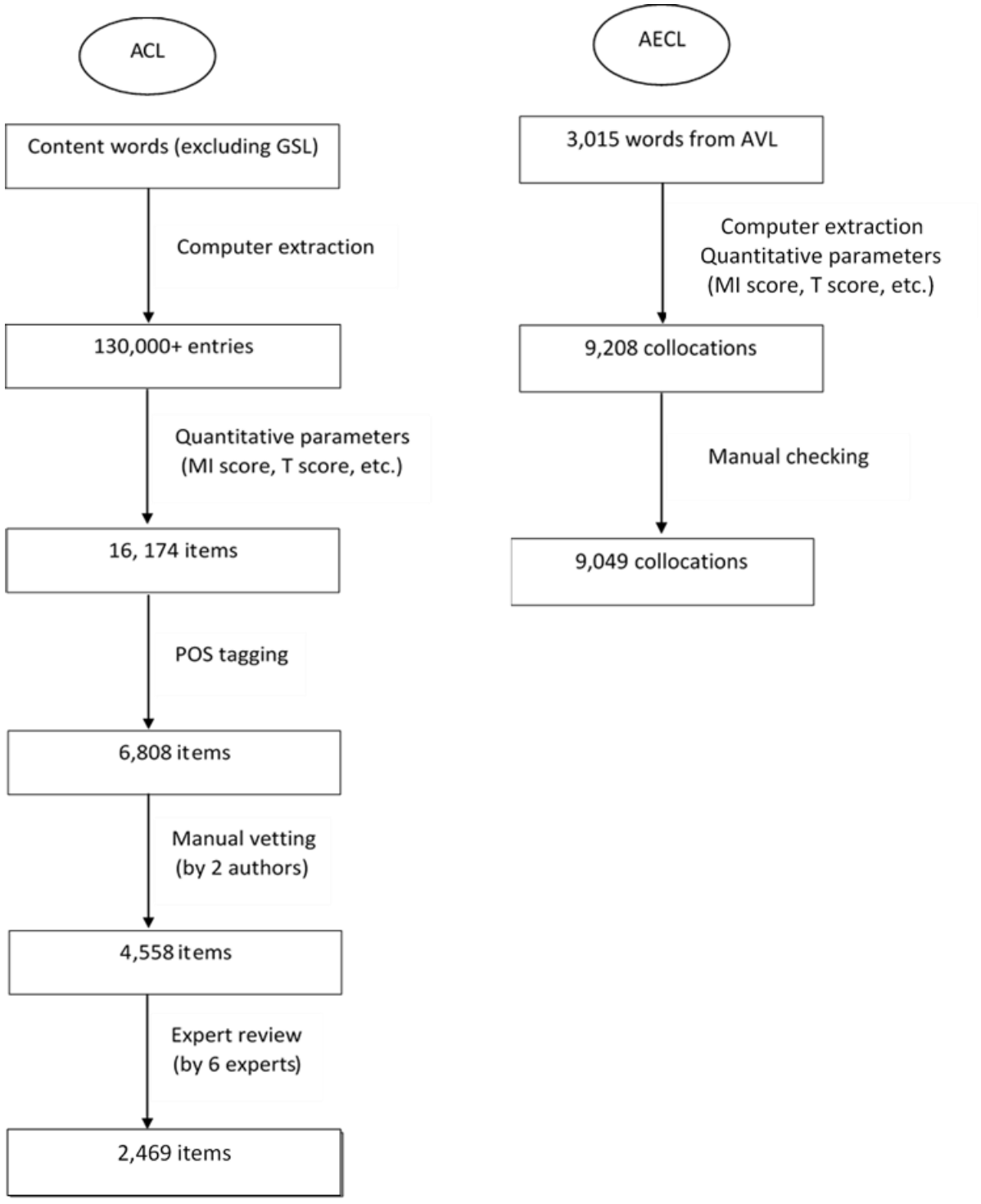
Finally, the statistical approaches of the list builders were similar, as both included MI score, t score, frequency and dispersion. Table 4.4 shows that both lists share the same threshold of MI score (at least 3) and dispersion (occurring at least 0.2 per million in each of the discipline divisions). The differences in the thresholds of t-score and frequency may be due to the difference in the corpus size of the two lists.

G. Manual checking

The last step of developing a word list – manual checking – clearly reflects the difference in the nature of the ACL (Ackermann & Chen, 2013) as a pedagogically relevant list and the AECL (Lei & Liu, 2018) as a statistics-based list. Figure 4.2 illustrates how the number of collocation items was narrowed down through different stages of making each list.

Figure 4.2

Stages of Refining the ACL and AECL



The ACL was refined using human intuition based on the high levels of expertise of professors and lecturers in the fields of Linguistics and English Language, as well as a dictionary consultant and lexicographer/ publisher. After steps involving computer extraction and applying statistical measures, 6,808 collocations were identified. The process of manual checking excluded 63.7% of those items to only retain collocations that are appropriate for teaching and learning purposes. The AECL, on the other hand, relied mostly on statistical measures and removed only 1.73% of the items from those identified by the computer extraction (proper noun-related and duplicates) as a final step for refining the list. Including almost every item from the computer extraction enabled Lei and Liu (2018) to avoid a possible source of selection bias in the study. However, this process resulted in a very long list which may discourage users.

H. Ordering items

Frequency is one of the most important guiding principles for usefulness when ordering word list items (Nation, 2013; 2016; Vilkaitė-Lozdiene & Schmitt, 2019). However, neither the ACL nor the AECL is published with frequency information nor are they ordered by frequency. Listing items alphabetically as the ACL (Ackermann & Chen, 2013) does could lead to form-based inference caused by words that look alike (Nation, 2000). For example, these items are listed near to each other in the ACL and learners may mismatch one for another: *future study – further study* or *racial difference – radical difference*.

Figure 4.3

Example of Item Listing in the AECL (Lei & Liu, 2018, p.239)

widely, adv; widely + adj; available
widely, adv; widely + v; accept, acknowledge, adopt, apply, differ, discuss, disperse, distribute, employ, know, perceive, recognise, regard, report, share, spread, use, vary

The AECL (Lei & Liu, 2018) has collocations listed as they would be in a dictionary (see Figure 4.3). This means that users can easily find the items. However, from the pedagogical view, it creates a lexical set with members having related meanings. Research has shown that presenting words with related meanings can make learning more difficult (Erten & Tekin, 2008; Nation, 2000). Ordering by

frequency might also be helpful, as other multiword list researchers (e.g., Durrant, 2009; Shin & Nation, 2008) have done. This is because frequency helps learners and teachers to immediately recognise which items are the most frequent in English and prioritise learning.

Table 4.5

Validation, Self-Criticism and Availability of the ACL and the AECL

Word list	The ACL (Ackermann & Chen, 2013)	The AECL (Lei & Liu, 2018)
I. Validation	Corpus-based comparison	Comparison with a collocation dictionary
J. Self-criticism	The use of human intervention in the manual checking stage might have removed some useful combinations from the final ACL.	The corpus was not completely balanced across discipline divisions as it was skewed towards social science and humanities.
K. Availability	Available on Pearson's website	Not publicly available

I. Validation

Ackermann and Chen (2013) and Lei and Liu (2013) employed different methods for validating their lists. The ACL (Ackermann & Chen, 2013) was validated by comparing the overall coverage of the list in the source corpus of academic texts and its coverage in a general corpus of a similar size. Their results show that the ACL has 14 times higher coverage in the academic corpus than in the general corpus, suggesting the importance of these collocations in an EAP context. Validating a list using corpus-based comparison seems to be standard practice (Miller & Biber, 2015). The validation of the AECL (Lei & Liu, 2018) was carried out through a comparison with a general English collocation dictionary. The resulting modest overlap of about 20% between the list and the dictionary suggests that the AECL mainly contains academic collocations rather than general collocations. This result can be partly explained by the fact that the AECL and the dictionary that was used to validate the list focus on different kinds of collocations. For example, verb + preposition and adjective + preposition collocations appear in the AECL but not the dictionary. Dictionary-based validation can be much more time-consuming than corpus comparison, and the dictionary used as a reference needs to be carefully considered.

For validation of a word list, Nation (2016) suggests comparing the coverage and lexical constituents against a competing list (see Chapter 2, Section 2.3.2). However, neither Ackermann and Chen (2013) nor Lei and Liu (2013) followed this step. Therefore, Sections 4.2.2 and 4.2.3 report the results of an assessment of overlap and the coverage of the ACL and the AECL.

J. Self-criticism

Ackermann and Chen (2013) and Lei and Liu (2013) acknowledge some limitations of their lists, including those related to unbalanced discipline division in the AECL (Lei & Liu, 2018) and extensive use of human intuition which led to the removal of some possible useful collocations in the ACL (Ackermann & Chen, 2013). That said, Ackermann and Chen (2013) argue that,

> existing corpus-driven multiword lists often fail to provide immediately usable resources for language learning, and it is only with expert intervention that raw data can be filtered and refined in order to extract the most informative and meaningful entries (p.246).

Read (2000) points out that statistical measures are useful tools for identifying multi-word units, but human intervention is still needed to select items that fit the purpose of the research.

K. Availability

While the ACL (Ackermann & Chen, 2013) is publicly available, the AECL (Lei & Liu, 2018) is not. This limitation of the AECL prevents the widespread use of the list.

Overall, with the application of Nation's (2016) adapted framework, the similarities and differences between the ACL and the AECL have been highlighted. The next section looks more closely into the lexical constituents between the lists to find out the overlap between them.

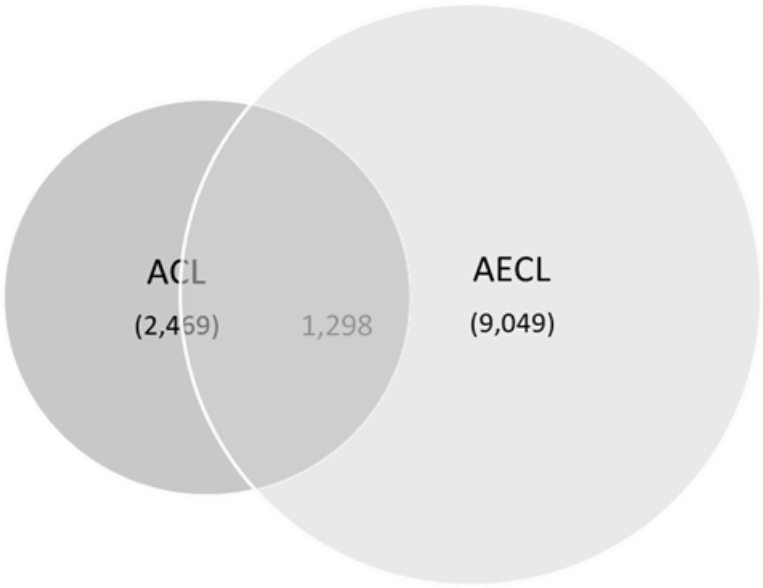
4.2.2 Lexical constituent overlap between the ACL and the AECL

The ACL and the AECL share 1,298 items (see Figure 4.4). Some examples of overlapping items are *mental health, physical activity* and *statistically significant*. More than half (52.57%) of the items in the ACL are also in the overlap list, but only a small part (14.34%, i.e., one-seventh) of the AECL overlaps with the ACL. The percentage of overlap indicates that although the two lists share the same

pedagogical purposes, they are quite different in lexical constituents. Noting overlaps between word lists is important because it tells us more about the nature of the word lists and possible differences between them.

Figure 4.4

Overlap Between the ACL and the AECL



There are four possible reasons for the differences between the lists. First, different procedures and criteria result in different lists. Ackermann and Chen (2013) gradually removed items after every step of making their ACL, while Lei and Liu (2018) retained almost all items from the computer extraction (Figure 4.2). Second, the corpora in the two studies are different in size, varieties of English, and text types. The corpus used to develop the AECL was nearly twice as large as the size of the ACL corpus. The ACL corpus includes American, Australian, British, Canadian and New Zealand English, while the AECL covers American, British and other English varieties which are not specifically mentioned in the published article. The AECL contains students' writing which is not included in the ACL corpus. Third, the kinds of collocations included in the ACL and the AECL also contribute to the difference between the two lists. That is, the ACL (Ackermann & Chen, 2013) contains only lexical collocations and the AECL (Lei & Liu, 2018) comprises both lexical and grammatical collocations. Another reason for this substantial difference is that other studies which

have compared multiword word units have found relatively low rates of overlap. Coxhead and Dang (2019) also found a modest overlap (about 20%) between Coxhead et al.'s (2017) list of multiword units in tutorials and laboratories, Biber et al.'s (2004) lexical bundles, and Simpson-Vlach and Ellis's (2010) spoken academic formulas.

4.2.3 Coverage of the ACL and the AECL over the COCA Academic and Fiction corpora

The overall coverage of the ACL and the AECL were compared over the COCA Academic corpus of 111 million words and the COCA Fiction corpus of 112 million words (Davies, 2012). As can be seen from the third column of Table 4.6, the overall coverage of both the ACL and the AECL is very low (<2.8 %). The results are in line with previous studies which also reported modest coverage percentages of multiword unit lists. For instance, Ackermann and Chen (2013) found that the ACL accounts for around 1.4% of an academic corpus of 25.6 million words. Another example comes from Miller (2020) whose idiom list makes up about 0.1% of a spoken academic corpus of nearly 1.7 million words. Using the same COCA academic corpus as the present study, Gardner and Davies (2014) found that the Academic Vocabulary List of single words covers almost 14% of the corpus. The findings of this study reflect the nature of multiword units which occur with much lower frequency compared to individual words.

Table 4.6

Coverage of the ACL and the AECL over COCA Academic and Fiction Corpora (%)

List	Number of items	Overall coverage over COCA Academic	Overall coverage over COCA Fiction	Average coverage over COCA Academic	Average coverage over COCA Fiction
ACL	2,469	0.84	0.06	0.000344	0.000020
Total AECL	9,049	2.76	0.22	0.000306	0.000024
Lex AECL	8,770	1.46	0.10	0.000166	0.000012
Gram AECL	279	1.30	0.12	0.004674	0.000430

It is not surprising that the overall coverage over the COCA Academic corpus increases along with the increase in the number of items in the lists (as demonstrated in Table 4.6, column 3). The overall coverage over the academic corpus of the AECL, with a significantly higher number of items (nearly

four times more than the ACL), is more than three times higher than that of the ACL. Strikingly, the Gram AECL with only 279 items accounts for almost half of the overall coverage of the total AECL. This finding echoes Durrant's (2009) finding that grammatical collocations are more frequent than lexical collocations. When comparing lexical collocations only, the overall coverage of the Lex AECL is just nearly twice the ACL despite being 3.5 times larger. This result is expected given the Zipfian nature of word frequencies, that is, as more words are included to the list, the extra coverage of the added items drops (Zipf, 1949). The overall coverage findings, therefore, do not imply that the ACL is superior to the AECL, but a further comparison is needed (see coverage comparison of the most frequent lexical items below).

The overall coverage of the two lists on the COCA Fiction corpus (see Table 4.6, column 4) also reveals interesting points. The coverage of the AECL over the academic corpus is nearly 13 times higher than its coverage over the fiction corpus. The ACL coverage is 17 times higher over the academic corpus compared to fiction. When carrying out their own validation of the ACL, Ackermann and Chen (2013) found that the coverage of their list was 14 times higher overall over the academic corpus than the general English corpus of similar size. Taken together, these results confirm that the ACL seems to perform well over different academic corpora and is more academic than general in nature.

If the average coverage metric over the academic corpus was applied, the ACL would provide a higher average coverage than the total AECL (as shown in Table 4.6, column 5). If we removed 279 grammatical collocations from the AECL, then the average coverage of the list would be 0.000166%, which is less than half of the ACL's average coverage (0.000344%). The average coverage figures over the fiction corpus (Table 4.6, column 6) reveal that the coverage of the Gram AECL is much higher than the other lexical lists (whose coverage figures are not much different from each other). However, findings of the average coverage of the lists are problematic given the frequency of words following Zipf's law (Zipf, 1949). Therefore, the lists were further broken down into equal sets of items for comparison. As explained in Chapter 3 (Section 3.1.1), for a more comprehensible comparison, the grammatical collocations in the AECL were removed and only lexical items (i.e., Lex AECL) were compared with those in the ACL.

Results from the coverage of the most frequent 500 lexical items (Table 4.7, column 4) indicate that these collocations are important because this group of a small number of items accounts for the majority of the overall coverage of the lists (64.29% of the ACL and 52.05 % of the AECL). When adding 500 more items (Table 4.7, column 5), the coverage of the ACL (Ackermann & Chen, 2013) increases by nearly 20% (from 0.54 to 0.70) while the coverage of the Lex AECL (Lei & Liu, 2018) increases by less than 14% (from 0.76 to 0.96). The coverage of 2,469 most frequent items (Table 4.7, column 6) shows that these items cover more than 86% of the overall coverage of the Lex AECL. These data from the AECL question the necessity of such a lengthy list when a shortened list would probably be more useful and practical for its intended teaching purpose.

Table 4.7

Coverage of the Most Frequent Items of the ACL and the Lex AECL over COCA Academic (%)

List	Number of items	Overall coverage on COCA Academic	Coverage of the most frequent 500 items	Coverage of the most frequent 1,000 items	Coverage of the most frequent 2,469 items
ACL	2,469	0.84	0.54	0.70	0.84
Lex AECL	8,770	1.46	0.76	0.96	1.26

This section has provided a thorough assessment of the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018), combining different methods of evaluation. First, the adapted evaluation framework from Nation (2016) provided the overall picture of how the lists were developed and validated. The lexical constituent comparison then gave a closer look at how the lists are similar and different. Finally, the lexical coverage helped to investigate the academic nature of the lists. For testing purposes in this thesis, the ACL (Ackermann & Chen, 2013) – the shorter list seems to be a better choice with the items selected based on both statistical measures and expert review. The findings of the evaluation demonstrate that the ACL (Ackermann & Chen, 2013) serves as a useful pedagogical list for its intended purpose, and it also provides high representativeness of academic collocations for testing purposes, especially when test practicality is taken into account. As a result, the ACL (Ackermann & Chen, 2013) was selected to develop the ACTs.

4.3 Sampling academic collocations from the ACL

The third step in the process of creating the ACTs was to sample the items from the list of Ackermann and Chen (2013). Not all of the 2,469 items in this list were used for the test development. Frequency criteria were used to narrow down the list. Frequency information of the collocations in COCA Academic and COCA Fiction was compared and only the collocations that were more frequent in the academic corpus were retained. This was to ensure that the selected items were the best representatives of academic collocations. Item frequency was calculated by adding all the frequency of different forms of an item together. Taking the academic collocation *differ significantly* as an example (see Table 4.8), the item frequency on COCA Academic is 1,047.

Table 4.8

Example of Item Frequency Calculation Using COCA Academic

Forms of the collocation	Frequency
differ significantly	610
differed significantly	346
differs significantly	80
differing significantly	11
Total frequency	1,047

Frequency information from the COCA corpus tool suggests that not all the items in the list of Ackermann and Chen (2013) are frequent in COCA Academic, as might be expected. The ACL item frequency in the COCA Academic corpus varies from 7472 to 3 occurrences per 112 million words. Some items are as frequent in the COCA Fiction corpus as in COCA Academic (e.g., *well aware, get involved, entirely different*, etc.), or even more frequent in the fiction corpus (e.g., *make contact, closer look, make (a) living*, etc.). The examples of items that are more frequent in the fiction corpus than the academic corpus indicate that some items in the list are possibly more general than academic collocations.

The selection of collocations for developing the ACTs was based on two principles. First, as the frequent items are likely to be more important items than the lower frequency ones, only items with a frequency in COCA Academic from 112 or above were selected into the item pool for the test development. This cut-off point was based on Ackermann and Chen's (2013) frequency criterion of at least one per million (i.e., frequency $\geq$ 1 per million) when developing their list. Second, to ensure

that only academic collocations were selected, only items that were at least 50% more frequent in COCA Academic than COCA Fiction (i.e., a 1.5 ratio) were retained. Different ratios ranging from 1.2 to 2.0 were trialled, and 1.5 seemed to be the most suitable ratio when it kept academic items such as *become obvious* and *give (an) indication*; and removed items commonly found in everyday conversation such as *give impression* and *make contact*. The same ratio was also used by Gardner and Davies (2014) when they compared frequency information from an academic corpus and a general corpus to create their Academic Vocabulary List. Altogether, 1,699 academic collocations in the Ackermann and Chen's (2013) list were selected as the item pool for the test development.

The next stage of the item sampling was dividing the 1,699 selected collocations into ten bands based on their frequency. Each band has approximately 170 collocation items as shown in Table 4.9. The purpose of this frequency band division was to evenly select items from different frequency ranges for the test development.

Table 4.9

Frequency Bands of the Collocation Item Pool

Band	Number of items	Frequency range
1	170	807 – 7472
2	170	507 – 803
3	170	382 – 506
4	170	306 – 380
5	171	254 – 305
6	171	212 – 253
7	166	180 – 211
8	174	154 – 179
9	168	133 – 153
10	169	112 – 132

The cut-off points for each frequency band were based on the natural breaks in the data set. As illustrated in Table 4.10, items 509 and 510 share the same frequency of 382, while the frequency of *make (a) transition* (item 511) is 380. The drop in the frequency at item 511 marks the end of Band 3 with 170 collocation items. There are cases when many collocations in the same band share the same frequency. As a result, the exact number of 170 collocation items for each band is difficult to maintain because it depends on where the breaks in the frequency data occur.

Table 4.10

Example of a Break in the ACL Collocation Frequency Data in COCA Academic

Band	Item number	Collocation item	Frequency on COCA Academic
Band 3	507	radically different	384
	508	essential component	383
	509	political stability	382
End of Band 3	**510**	**specific type**	**382**
Band 4	**511**	**make (a) transition**	**380**
	512	numerous studies	379
	513	foreign investor	379
	514	fully understand	379

A total of 60 academic collocations were selected to develop 60 AC Recognition Test items and 60 AC Recall Test items. A random selection of six collocations from each frequency band was made using Randomise tool in Excel. This sampling rate (1:29) is approximately equivalent to the sampling rate of the Vocabulary Levels Test (1:33) (Schmitt et al., 2001). The random selection resulted in the number of test items for each collocation kind as shown in Table 4.11.

Table 4.11

Collocation Kind Ratio From a Random Selection of Test Items

Kinds of collocation	Number of test items
Adjective + noun	31
Noun + noun	1
Verb + noun	17
Verb + adjective	1
Adverb + verb	1
Verb + adverb	1
Adverb + verb participle	3
Adverb + adjective	5
Total	60

In addition to the frequency criteria, this study deliberately selected items so that all the kinds of collocations in the Ackermann and Chen (2013) list were included in the ACTs. Interestingly, there was no noticeable difference in the kind ratio between the original ACL (Ackermann & Chen, 2013) and the 1,699-item version (Table 4.12, columns 2 and 3). The number of test items for each collocation kind was based on the 1,699-item ACL kind ratio (Table 4.12, column 4).

Table 4.12

Collocation Kind Ratio and Number of Test Items for Each Kind

Kinds of collocation	Ratio in the ACL with 2,469 items	Ratio in the ACL with 1,699 items	Number of test items
Adjective + noun	71.8%	72.3%	42
Noun + noun	2.5%	3.1%	2
Verb + noun	12.6%	11.9%	7
Verb + adjective	1.2%	1.2%	1
Adverb + verb	0.6%	0.8%	1
Verb + adverb	1.2%	1.0%	1
Adverb + verb participle	5.0%	5.5%	3
Adverb + adjective	5.0%	4.2%	3
Total	100%	100%	60

In order to maintain the kind ratio as shown in Table 4.12, nine adverb + verb collocations and two adverb + adjective collocations in Table 4.11 were replaced by 11 random adjective + noun collocations in the equivalent frequency bands. The final goal was to create 60 test items for each of the two academic collocation tests. A further 20 collocations were selected for test items as a contingency measure. This meant that each collocation kind had at least one extra item (see Table 4.13). Because random selection was also employed until the collocation kind criterion was met, the frequency bands of the additional items were not controlled. After the piloting stage (see Section 4.6), the best 60 items were selected from 80 items to be used for the validation study.

Table 4.13

Collocation Kinds of Additional Items and the Total Number of Pilot Test Items

Kinds of collocation	Number of additional items	Total number of pilot items
Adjective + noun	9	51
Noun + noun	1	3
Verb + noun	3	10
Verb + adjective	1	2
Adverb + verb	1	2
Verb + adverb	2	3
Adverb + verb participle	1	4
Adverb + adjective	2	5
Total	20	80

4.4 Selecting the test formats

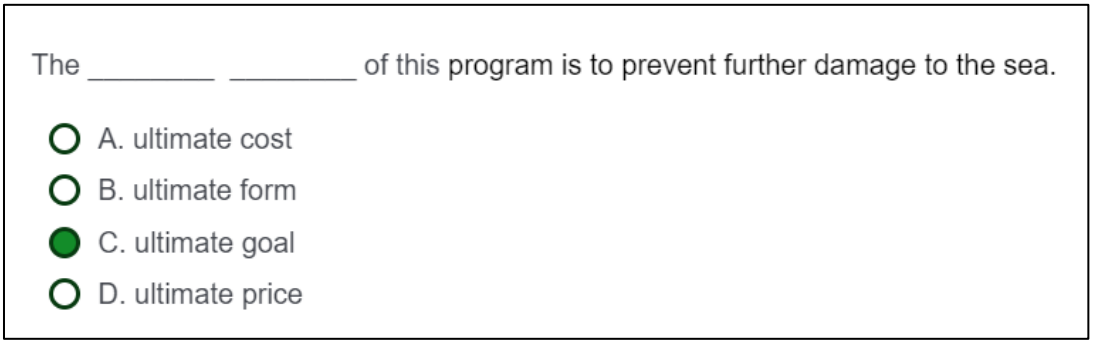
The fourth step of developing the ACTs was to select the formats for the AC Recognition Test and the AC Recall Test. The same academic collocations and sentence prompts were used in both tests. Targeting the same academic collocations can help to compare the recognition and recall knowledge elicited by the two tests. Moreover, as the tests were administered together, the use of the same sentence prompts in both tests might reduce the reading burden for test-takers. The formats of the AC Recognition Test and the AC Recall were selected following the review of literature in Chapter 2 (Section 2.4) with adaptations based on feedback from pilot participants. These test formats are in turn described as below.

4.4.1 The AC Recognition Test format

A multiple-choice format was selected for the AC Recognition Test because it has more advantages than other test formats such as Yes/ No or matching (see Chapter 2, Section 2.4.1). A sentence prompt was provided with two blanks, one for each word in the academic collocation being tested. The collocations were always presented as two-word collocations without any other words in between, except for an article in some cases (e.g., *address the issue*). Test-takers were asked to select one academic collocation to fill the blanks from four which were provided. The correct answers were distributed evenly among the four options A, B, C and D. Figure 4.5 gives an example of an AC Recognition Test item.

Figure 4.5

Example of an AC Recognition Test Item



Four options were used instead of three to reduce the chance of answering correctly by random guessing. According to Nation and Webb (2011), the quality of distractors is more important than the number of distractors, but for a vocabulary test, a four-option format is recommended. An 'I don't know' option was not provided to avoid the issues related to analysing and interpreting this response. Nation and Webb (2011) (also see Zhang, 2013 and Stoeckel, Bennett & McLean, 2016) argued that test-takers with different personalities make use of an 'I don't know' option in different ways. For example, they may not want to take a risk and choose this option even when they may have partial knowledge of the words tested. Some test-takers may choose 'I don't know' so they can skip a test item. Some test-takers who really do not know an answer may want to take a guess instead of choosing an 'I don't know' option. Scoring the 'I don't know' option is also difficult. If it is counted as a wrong answer, risk-takers benefit more than those who honestly reflect their knowledge. 'I don't know' should not be counted as a missing response either because test-takers may choose it on purpose, meaning that test scores are not deducted even when test-takers do not have the knowledge of the item being tested.

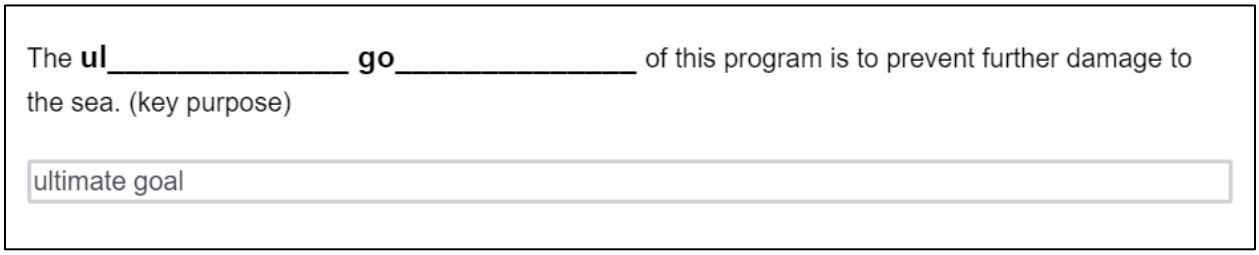
The ACTs were administered on Qualtrics (https://www.qualtrics.com/) as a testing platform. The function of *Request response* on Qualtrics let test-takers know how many questions were unanswered on the page but they could choose to continue without answering. This means that all the test items were delivered to students, and they had the chance to give their answers. Therefore, any blank answer was treated as a wrong answer in this study.

4.4.2 The AC Recall Test format

A gap-filling format was selected for the AC Recall Test, because it is a commonly used recall task to measure a test-taker's ability to produce a suitable collocation for a provided context (see Chapter 2, Section 2.4.2). In this test, two initial letters of each word in a collocation were provided and the meaning of the collocation was given in brackets at the end of the sentence prompt. Test-takers were requested to provide an academic collocation in a box below each question. An example of an AC Recall Test item is illustrated in Figure 4.6. This test also employed the web-based format using Qualtrics.

Figure 4.6

Example of an AC Recall Test Item



This format of the test bears some resemblance to the test of Fernández and Schmitt (2015) (see Chapter 2, Section 2.4.2.3), but the AC Recall Test has two different features. First, unlike Fernández and Schmitt (2015) who provided just one first letter of each word in a collocation pair, the AC Recall Test consistently used two initial letters. Second, the AC Recall Test replaced L1 context sentences as in the test of Fernández and Schmitt (2015) with English meanings of collocations in brackets at the end of sentence prompts (e.g., *key purpose* in Figure 4.6). This is because the participants in the present study came from a variety of first language backgrounds. Informal piloting with three native speakers and three non-native English speakers who were researchers and PhD students in applied linguistics showed that two initial letters and meanings of collocations were necessary in order to help ensure that the target collocations were elicited as much as possible. Although this format may not be ultimate for measuring recall knowledge (see Chapter 8, Section 8.3), no better alternative could be found for the testing purposes of this study.

4.5 Writing the test items

The fifth step in the test development process involves the creation of sentence prompts, provision of collocation meanings and selection of distractors, as in turn described below.

4.5.1 Creation of sentence prompts

Creating sentence prompts which are comprehensible for the test takers is an important step in vocabulary test development. One option was to use authentic academic texts taken from concordance lines in the COCA Academic corpus, but it proved to be too challenging because the concordances were usually discipline-specific. Test-takers without background knowledge in the field

would find it hard to understand the context, which might prevent them from choosing or providing the correct academic collocations. For example, the following context sentence was found in the COCA concordances for the academic collocation *ultimate goal*:

> The **ultimate goal** of this revised model was to foster documentation of curricular modifications in situations where deficiencies existed or areas for improvement became evident (Balotsky et al., 2016, p.77).

In the above example, with knowledge of the first 2,000 word families in the BNC/COCA lists (Nation, 2012), learners can only reach 68% comprehension of the concordance. Learners need a vocabulary size of at least 4,000 words to understand the whole sentence. This is in line with Ballance and Coxhead (2020) who found that concordances from authentic corpora have an average vocabulary load of 4,000–5,000 word families. The option of using concordances was abandoned and the process of considering how to create sentence prompts concluded with the development of the following criteria:

1. The sentences should create a suitable context in which the target collocation fits in.
2. The sentences should not be too long.
3. The sentences should be formal and impersonal to reflect the academic register.
4. The sentence prompts should avoid complicated structure.
5. The vocabulary used to write the sentence prompts should be restricted to words from the first 2000 word families of the BNC/COCA word family lists (Nation, 2012). The assumption is that students who would like to start an academic study at an English-medium university should know these words.

Following the criteria, I developed the sentence prompts on my own. There were a few sentences using definitions in online dictionaries which were used as possible models, but the language in them needed to be simplified. For example, the following context sentence was found in https://examples.yourdictionary.com/ for the collocation *social norms*:

> **Social norms**, or mores, are the unwritten rules of behavior that are considered acceptable in a group or society.

The example sentence above might be difficult for learners because the words “mores” and “behavior” used in the sentence belong to the 10,000 and 3,000 word family levels of the BNC/COCA lists (Nation, 2012), respectively. Consequently, the sentence was slightly modified to be used in the tests as follows:

> An unwritten rule of manners that are considered acceptable in a group or society is called a _____ _____.

A total of 80 sentence prompts were developed for being used in both the AC Recognition Test and the AC Recall Test. The shortest sentence has five words and the longest has 27 words. The average length of a sentence prompt is 12 words. Only four verb tenses were used to write the sentences, including the simple present, simple past, present perfect and the simple future. These are the most commonly used verb tenses in English. All the sentence prompts were checked against the online vocabulary profile tool at https://www.lextutor.ca/ to ensure that the vocabulary used belongs to the most frequent 2,000 word families of the BNC/COCA lists (Nation, 2012). Two native speakers who are researchers in applied linguistics were also invited to check whether the sentence prompts were clear, simple and natural, and if they provided enough context to elicit the target academic collocations.

4.5.2 Provision of collocation meanings

The meaning of the target collocations was expressed by a short phrase provided in brackets at the end of the sentence prompts in the AC Recall Test. The length of the phrases ranges from one to ten words. The average length is three words. All of the meaning phrases were also checked at https://www.lextutor.ca/ to ensure that all the words are from the most frequent 2,000 word families of the BNC/COCA lists (Nation, 2012) and were reviewed by the two researchers mentioned in Section 4.5.1.

4.5.3 Selection of distractors

The AC Recognition Test used a multiple-choice format with four options provided. Apart from one academic collocation as the correct answer, the other three distractors followed three principles below:

1. All distractors are academic collocations found in COCA Academic with mutual information scores above three (i.e., MI score ≥ 3.0). This is the threshold where a word combination is called a collocation (Church & Hanks, 1990).
2. All the options fit the context grammatically.
3. The meaning of the distractors is not plausible in the given context.

To search for the distractors, the COCA corpus tool was employed as illustrated in Figure 4.7. One of the words in the two-word collocation was randomly used as a node word, and the part of speech of the collocate was selected. The searching span was one word either to the left or to the right of the node word. The corpus was selected as Academic only. The MI score was set at three as the minimum value to ensure that all the results returned were academic collocations.

Figure 4.7

The COCA Corpus Tool for Distractor Search

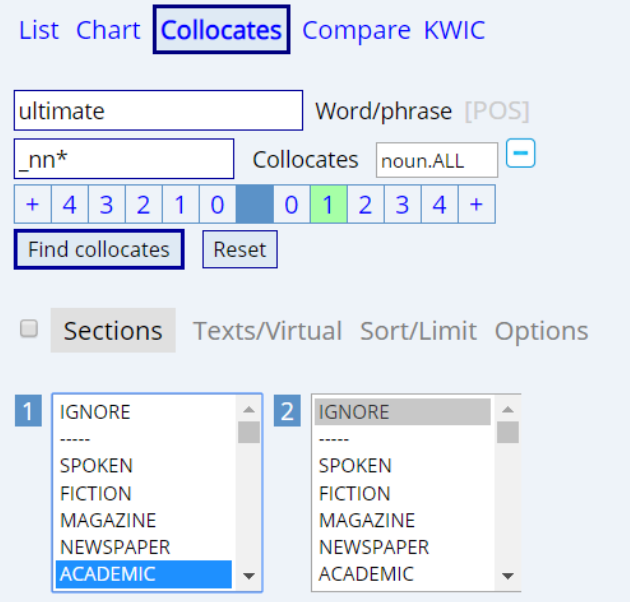
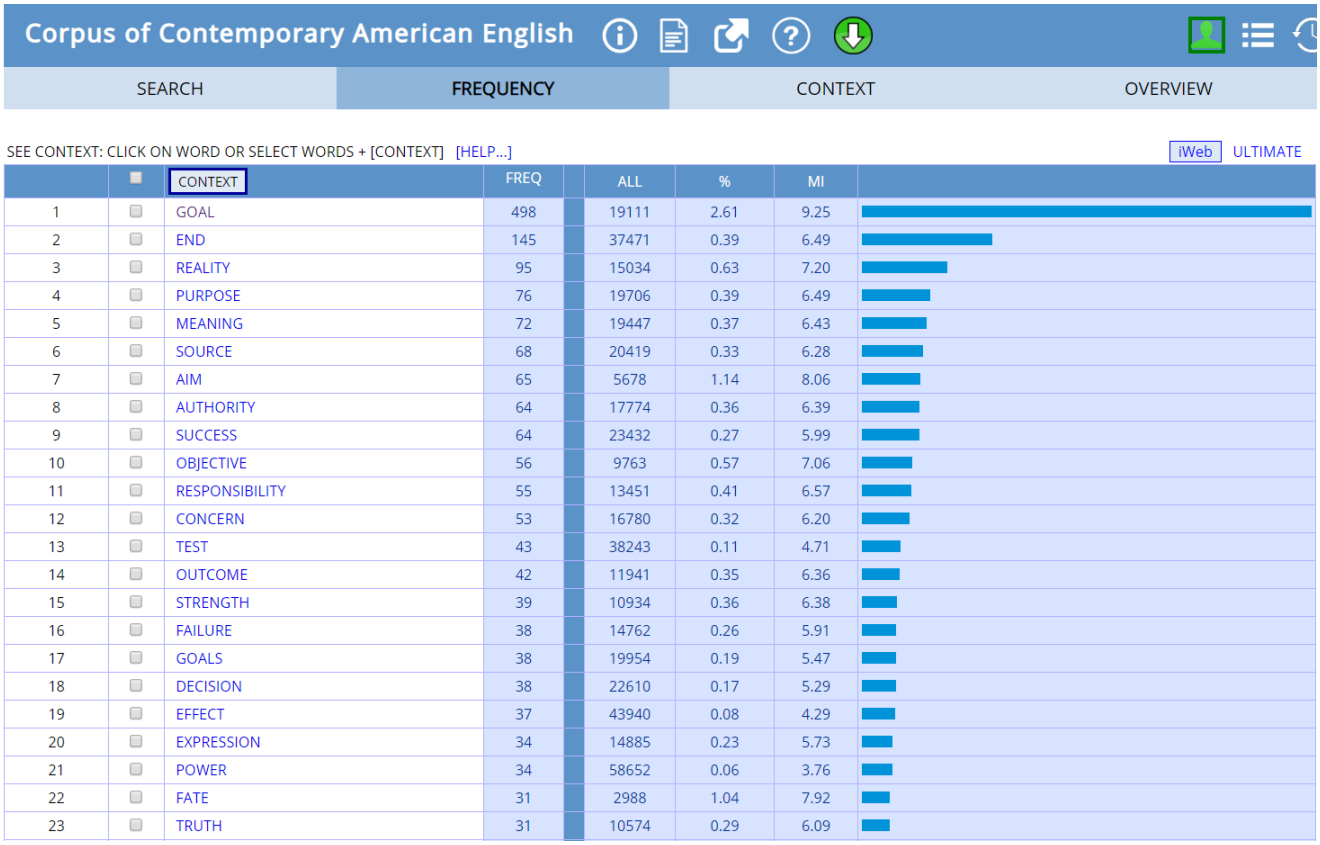


Figure 4.8 gives an example of searching results for two-word academic collocations of the word "ultimate" + a noun.

Figure 4.8

Example of COCA Search Results



	CONTEXT	FREQ	ALL	%	MI
1	GOAL	498	19111	2.61	9.25
2	END	145	37471	0.39	6.49
3	REALITY	95	15034	0.63	7.20
4	PURPOSE	76	19706	0.39	6.49
5	MEANING	72	19447	0.37	6.43
6	SOURCE	68	20419	0.33	6.28
7	AIM	65	5678	1.14	8.06
8	AUTHORITY	64	17774	0.36	6.39
9	SUCCESS	64	23432	0.27	5.99
10	OBJECTIVE	56	9763	0.57	7.06
11	RESPONSIBILITY	55	13451	0.41	6.57
12	CONCERN	53	16780	0.32	6.20
13	TEST	43	38243	0.11	4.71
14	OUTCOME	42	11941	0.35	6.36
15	STRENGTH	39	10934	0.36	6.38
16	FAILURE	38	14762	0.26	5.91
17	GOALS	38	19954	0.19	5.47
18	DECISION	38	22610	0.17	5.29
19	EFFECT	37	43940	0.08	4.29
20	EXPRESSION	34	14885	0.23	5.73
21	POWER	34	58652	0.06	3.76
22	FATE	31	2988	1.04	7.92
23	TRUTH	31	10574	0.29	6.09

In a few cases, only one or two collocations qualified as distractors, and the third distractor could not be found. The search was restarted but the word which had been used as node word was swapped to the other word. To ensure that all the distractors were semantically implausible in the given contexts, they were reviewed by the two linguists mentioned in Section 4.5.1.

4.6 Piloting and finalising the ACTs

The final step of developing the ACTs was to pilot and finalise the tests. The main purposes were to:

- pilot two scoring methods for the AC Recall Test: strict scoring and lenient scoring
- pilot the scoring procedure for the AC Recall Test
- test whether the ACTs conformed to the Rasch model
- choose the best 60 out of 80 test items for the ACTs

In the next section, the scoring methods for the AC Recognition Test and the AC Recall Test are detailed. The procedure for removing test items using Rasch analysis is also described.

4.6.1 Scoring the ACTs

In the first step, the pilot AC Recognition Test and the pilot AC Recall Test were scored separately. For each test item in the AC Recognition Test, there was only one correct answer among four options. Test-takers received one point for the right choice and a zero point for the wrong choice (i.e., correct = 1, incorrect = 0). The AC Recall test was piloted with two different scoring methods: strict scoring and lenient scoring. This was done to determine which method would be more appropriate for the main study. Under strict scoring, test-takers received one point for every correct academic collocation that exactly matched the answer key, and zero points for any other responses. The lenient scoring, on the other hand, did not take into account spelling and grammatical mistakes (as long as it was clear which word test-takers intended) and followed the scoring map presented in Figure 4.9. If test-takers produced a response that was different from the answer key, their answer could be accepted if it a) fitted the sentence prompt as well as the meaning provided, and b) was an academic collocation.

To decide whether a learner's response could be counted as a correct academic collocation or not, three different methods were applied as illustrated in Figure 4.10. First, the word combination produced by test-takers was checked against COCA Academic to see if its MI score was from three or above which is a threshold to decide whether a combination is a collocation (Church & Hanks, 1990). When a learner's response could not be found in COCA Academic or its MI score was under three, it was searched for in the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018). If the response still did not exist in either of the lists, expert judgement was employed as to whether that answer could be an accepted academic collocation for the provided context or not. The rater in this study was a native English speaker who is an established researcher in academic vocabulary. This rater also checked academic corpora and consulted other academic colleagues to reach her final decision. This scoring process was to make sure that participants' knowledge of academic collocations was not underestimated.

Figure 4.9

Lenient Scoring Map for the AC Recall Test

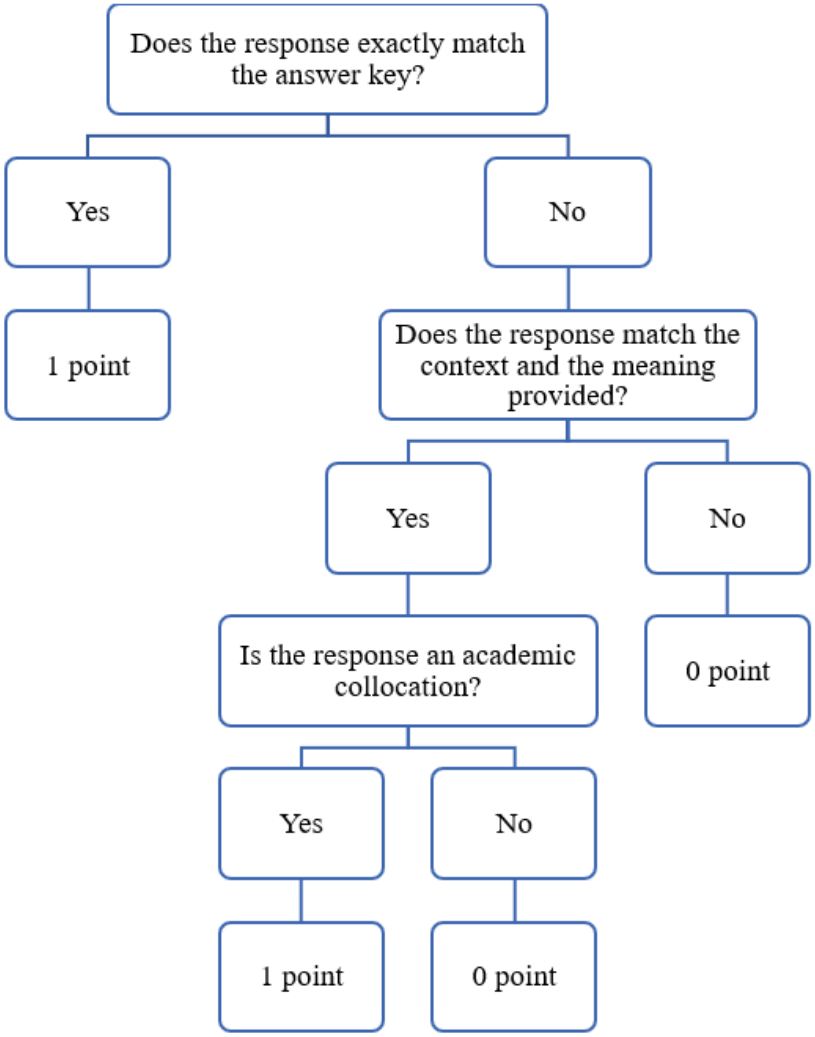
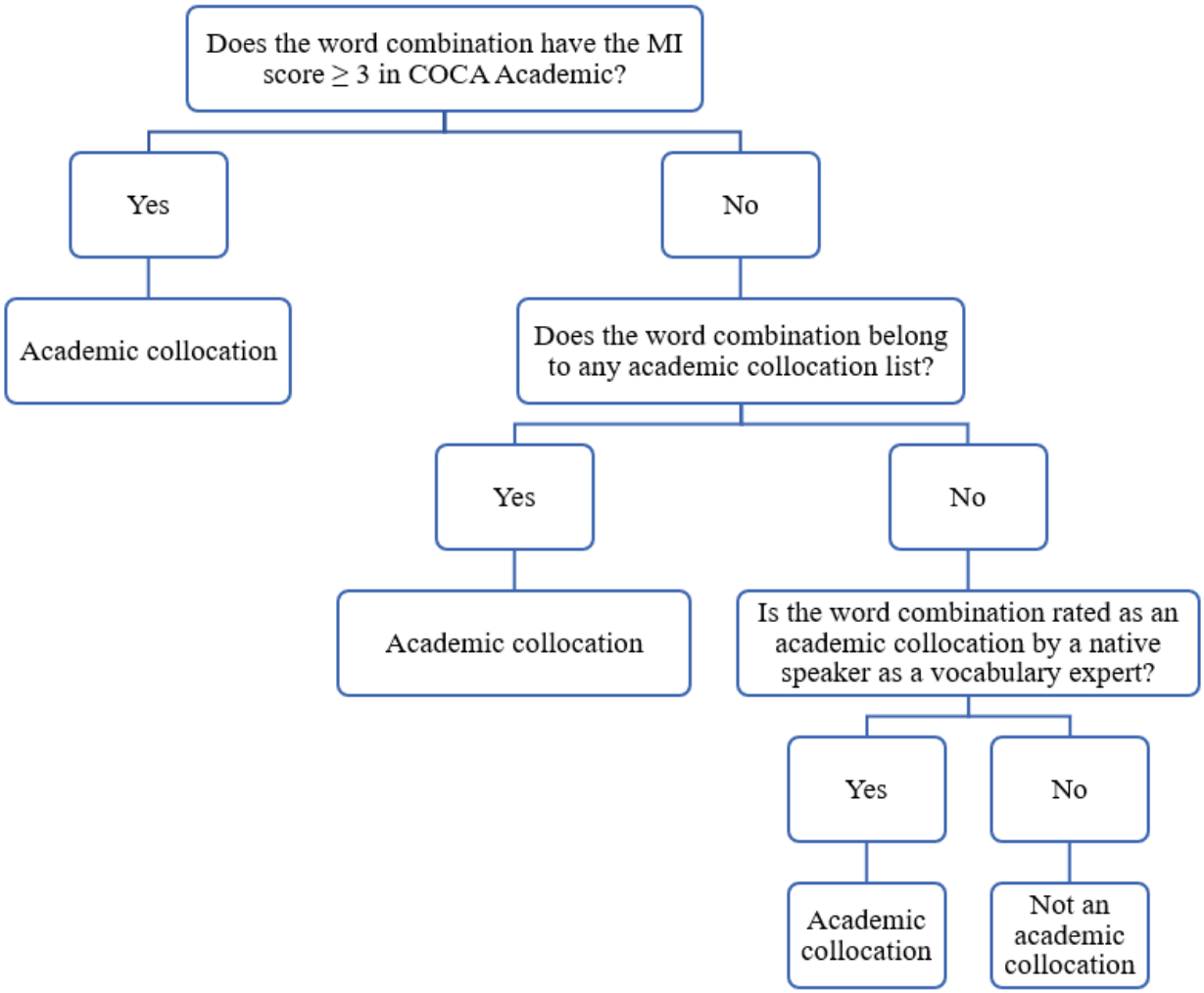


Figure 4.10

Procedures for Determining Whether a Response Is an Academic Collocation

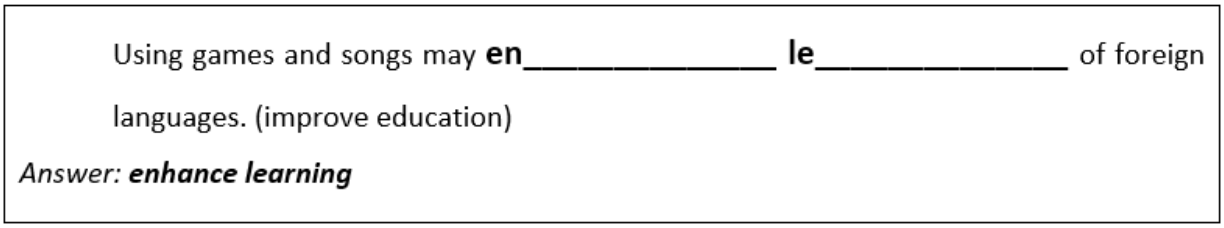


The scoring procedure was also a part of the piloting, and the process was decided on practicality. Firstly, COCA Academic – the largest available corpus of academic vocabulary – is simple to use and accessible to all researchers. That said, not every academic collocation will be represented in even a large corpus. Using published word lists is also an easy way to check the items as long as the lists are accessible. However, the number of items in a list is often limited. Finally, employing a rater is a good way to judge whether a response is an academic collocation because depending on different contexts, it might or might not be accepted and only humans can have that flexible interpretation. To have a reliable judgement, raters should be experienced readers of academic texts, and they can use

other resources to help decide on academic collocations. After checking against the corpus and the lists, the number of items that need human judgement was narrowed down, meaning the time and effort for raters was not demanding.

Figure 4.11

Example of an AC Recall Test Item



Let us take one test item as an example (see Figure 4.11) to illustrate lenient scoring (see Table 4.14 for the suggested scoring). The intended correct answer for this test item is ***enhance learning***. In lenient scoring, ***enhances learning*** and ***enhace learning*** are also accepted as correct answers because this test is not intended to measure learners' knowledge of grammar and spelling. Another acceptable answer is ***enrich learning*** which is also an academic collocation (MI score = 3.66 in COCA Academic) and fits the given context and the meaning in the brackets. On the other hand, ***encourage learning*** is an answer that fits the sentence prompt but not exactly the meaning in the brackets; therefore, it is marked as incorrect. Another response which is considered incorrect is ***enhancement learning***. The wrong use of part of speech makes the word combination become deviant and there is no such word combination in COCA Academic. Other answers, such as ***enforce learning*** or ***enlarge learning***, are not acceptable either because they did not survive the scoring procedure of checking the COCA Academic corpus, the two academic collocation lists and consulting the expert rater. So even though participants can produce one correct word of the two-word collocation, these answers only show that test-takers can understand the context and the given meaning, but they do not have knowledge of the academic collocation. As a result, only one correct word is treated the same as an incorrect answer.

Table 4.14

Example of Lenient Scoring Using Test Item in Figure 4.11

Answer	Score	Explanation
enhance learning	1	Response matches the answer key
enhances learning	1	Grammatical error is not counted
enhace learning	1	Spelling error is not counted
enrich learning	1	Response fits the context and the meaning in brackets
encourage learning	0	Response fits the context but not the meaning in brackets
enhancement learning	0	Word combination is not accepted in English
enforce learning	0	Word combination is not accepted in English
enlarge learning	0	Word combination is not accepted in English

To decide which scoring method was more suitable for the validation study, a paired sample t-test was conducted to compare the test results of the strict scoring and the lenient scoring. The results from the strict scoring (M = 39.50, SD = 14.90) and the lenient scoring (M = 46.76, SD = 14.95) indicate that the participants' test scores were significantly higher with the lenient scoring method, $t(37) = 14.09$, $p < .001$. The piloting also showed that the lenient scoring worked well with the AC Recall Test and better reflected participants' knowledge of academic collocations; hence, it was applied as the scoring method for the validation study.

4.6.2 Testing the Rasch model and finalising the ACTs

As discussed in Chapter 3 (Section 3.1.3), the Rasch model (Rasch, 1993) was used to investigate whether the test items performed well to measure the target construct. One of the main aims of the piloting was to choose 60 best items from 80 items to be used for the validation study. When an item was removed from the pilot AC Recognition Test, it was also removed from the pilot AC Recall Test. As a result, the piloting results of the two tests, which were administered to 79 students in Vietnam and New Zealand (see Section 3.1.3), were combined in order to select the best 60 items for both tests. The process of removing test items followed four steps:

Step 1	List all the items which were identified as misfits in the AC Recognition Test according to the Winsteps Rasch software analysis.
Step 2	List all the items which were identified as misfits in the AC Recall Test according to the Winsteps Rasch software analysis.
Step 3	Investigate the reasons causing misfits for the items in both tests and remove items with clear issues.
Step 4	Control removed items so that each frequency band contains only six items. At the same time, control the kind ratio so that the items removed did not change the ratio presented in Table 4.12.

Following the first two steps, all the underfit (MNSQ > 1.5) and overfit (MNSQ < 0.5) items of the AC Recognition Test and the AC Recall Test were listed for investigation. Table 4.15 shows the number of misfit items. A total of 37 items were listed as misfits from the Rasch analysis of the two tests, including five items misfitting in both tests.

Table 4.15

Misfit Items in the Pilot ACTs

	AC Recognition Test	AC Recall Test	Total
Underfit items	10	8	18
Overfit items	18	6	24
Total	28	14	37*

Note. *Five duplicate items were subtracted from the total.

Apart from items whose source of misfit could not be identified, three main issues were found to cause misfit after Step 3 (see Table 4.16). First, the wording was a factor in item misfits. Some sentence prompts were not transparent enough to provide a clear context for test-takers to answer the test items. Second, the Rasch analysis showed that some distractors in the AC Recognition Test did not work well, which also contributed to item misfit. Lastly, some items were misfits because some test-takers performed unexpectedly (i.e., misfit persons).

Table 4.16

Number of Misfit Items Under Possible Sources of Misfit

Source of misfit	AC Recognition Test	AC Recall Test	Total
Misfit person	2	2	4
Poor wording	1	4	4*
Poor distractor	7	0	7
Unidentified source	18	8	22**

Note. *One duplicate item was subtracted from the total, **Four duplicate items were subtracted from the total.

Four items (3, 15, 21, and 54) caused by misfit persons were retained because these items became fit after removing responses from these test-takers. A total of 11 misfit items were removed in this step because of unclear wording (items 47, 57, 59 and 77) and poor distractors (items 5, 11, 20, 29, 31, 43, and 68). There were 22 items (2, 7, 14, 16, 17, 22, 23 35, 36, 37, 45, 49, 51, 52, 53, 55, 69, 71, 72, 73, 76, 79) whose source of misfit could not be identified. The misfit could be the randomness predicted by the Rasch model.

Step 4 involves balancing removed items to meet the condition of collocation kind ratios and the number of items for each frequency band. After removing 11 items in Step 3, the number of collocations left is described in Table 4.17.

Table 4.17

Number of Collocations After Removing Items in Step 3

Collocation kinds	Number of collocations as planned	Number of collocations after Step 3	Number of collocations need removing
Adjective + noun	42	44	2
Noun + noun	2	3	1
Verb + noun	7	8	1
Verb + adjective	1	2	1
Adverb + verb	1	2	1
Verb + adverb	1	1	0
Adverb + verb participle	3	4	1
Adverb + adjective	3	5	2
Total	60	69	9

Nine more items were removed in Step 4 to reach the final aim of 60-item tests. Three items (9, 67 and 75) were deleted because of the kind limit and because their repeated initials made the answers obvious when eliciting the same word whenever they appeared. For example, with the initial "***eq_____***" (e.g., *equal opportunity*), test-takers always produced the response "*equal*". Another six items (8, 23, 37, 40, 49 and 53) were removed to meet the requirement of the number of collocations for each collocation kind described in Table 4.12 and to limit the number of collocations for each frequency band to six.

After the two tests had been finalised with 60 selected items, a Rasch analysis was run again for the 60-item test versions of the pilot ACTs. The summary statistics of the pilot test scores are presented in Table 4.18. The perfect score for each test was 60.

Table 4.18

Summary Statistics of Pilot Test Scores

Pilot test scores	AC Recognition Test (N = 42)	AC Recall Test (N = 38)
Mean	43.6	34.4
SD	13.0	12.2
Max	58	57
Min	15	10

The reliability indices of 60-item test versions of the pilot ACTs were checked to confirm whether the data conform to the Rasch model or not (see Table 4.19). Winsteps Rasch software provides two reliability indices, namely person reliability and item reliability. Two separation indices (i.e., person separation and item separation) give an additional evaluation of the instruments.

Table 4.19

Reliability Indices of 60-Item Test Versions of the Pilot ACTs

Pilot test scores	AC Recognition Test	AC Recall Test
Person reliability	.92	.93
Person separation	3.31	3.51
Item reliability	.83	.91
Item separation	2.24	3.10

Person reliability depends on the test length: the longer the test, the higher the person reliability (Linacre, 2019). Person reliability above .90 indicates that the tests can discriminate the person sample into three or four levels (Linacre, 2019). A person separation index above 3.00 represents an excellent level of separation (Fisher, 1992). The person separation indices of the two tests show that both tests were sensitive enough to distinguish between high and low performers. Item reliability depends on person sample size, which means the larger sample, the higher reliability (Linacre, 2019). The item reliability indices (above 0.90) and the item separation indices (above 3.00) suggest that the person sample is large enough to confirm the item difficulty hierarchy of the instrument (Linacre, 2019). Overall, both the AC Recognition Test and the AC Recall Test produced good reliability indices, but a larger person sample size was needed to confirm the item reliability of the AC Recognition Test. Consequently, the 60-item versions of the AC Recognition Test and the AC Recall Test (see Appendices C and D) were administered to bigger groups of participants to seek validity evidence (see Chapters 5 and 6 for the findings).

4.7 Chapter summary

This chapter clarified the six steps involved in developing the ACTs. The development process started with the evaluation of the two academic collocation lists and the results indicated that the Academic Collocation List (Ackermann & Chen, 2013) worked better for the testing purposes in the present study. The chapter provided a detailed description of item selection from the list of Ackermann and Chen (2013), justification for the test format and the process for writing the test items. This chapter also presented the pilot results which indicated that the ACTs worked well as intended and were ready for the validation study. The process of developing academic collocation tests from a corpus-based word list will be further discussed in Chapter 7 (Section 7.2) in terms of opportunities and challenges. Following Chapter 4, the next two chapters report the validation results of the ACTs, with Chapter 5 focusing on the Evaluation inference and Chapter 6 on the Generalisation and Extrapolation inferences in the argument-based validation framework of the ACTs.

Chapter 5 Validation results: Evaluation inference

This chapter reports the findings that served as evidence for the assessment of the Evaluation inference in the argument-based validation framework applied for the Academic Collocation Tests (ACTs). The Evaluation inference states that test-takers' performance on the ACTs was appropriately observed and scored. This inference relies on warrants concerning the appropriateness of the test characteristics, the testing condition and the scoring method. The evidence reported in this chapter helps to assess the Evaluation inference and the findings are presented in relation to the following research questions (RQ):

RQ1. To what extent the items on the ACTs are appropriate to measure the intended construct of academic collocations?
RQ2. Does the testing condition allow test-takers to demonstrate their knowledge of academic collocations?

Chapter 5 begins with quantitative results. Sections 5.1 and 5.2 present descriptive statistics of the test scores and item analysis based on the Rasch model (Rasch, 1993). The next three sections report qualitative evidence, including test-takers' opinions about the test formats (Section 5.3), their reflections on the test-taking strategies (Section 5.4) and the test-taking processes (Section 5.5). Section 5.6 then provides an overall assessment of the Evaluation inference before Section 5.7 summarises the findings of this chapter.

5.1 Descriptive statistics of the test results

This section provides an overall picture of how the participants scored on each test in this research, which forms the basis for further statistical analysis. Results of the ACTs and the Vocabulary Size Test (VST) (Nation & Beglar, 2007) are summarised in Table 5.1. The mean score of the AC Recognition Test was 44.57 out of 60 (74.28%) with a standard deviation of 13.80. The mean score of the AC Recall Test was 26.03 out of 60 (43.38%) with a standard deviation of 14.82. The participants, therefore, had a higher average score on the AC Recognition Test than the AC Recall Test. The mean score of the VST was 95.91 out of 140, meaning that on average the participants had a vocabulary size of around 9,600 word families, with a standard deviation of 23.58. The score data

on the three tests are not normally distributed as indicated by the significant *p* values (<.001) of the Shapiro-Wilk test (last column in Table 5.1).

Table 5.1

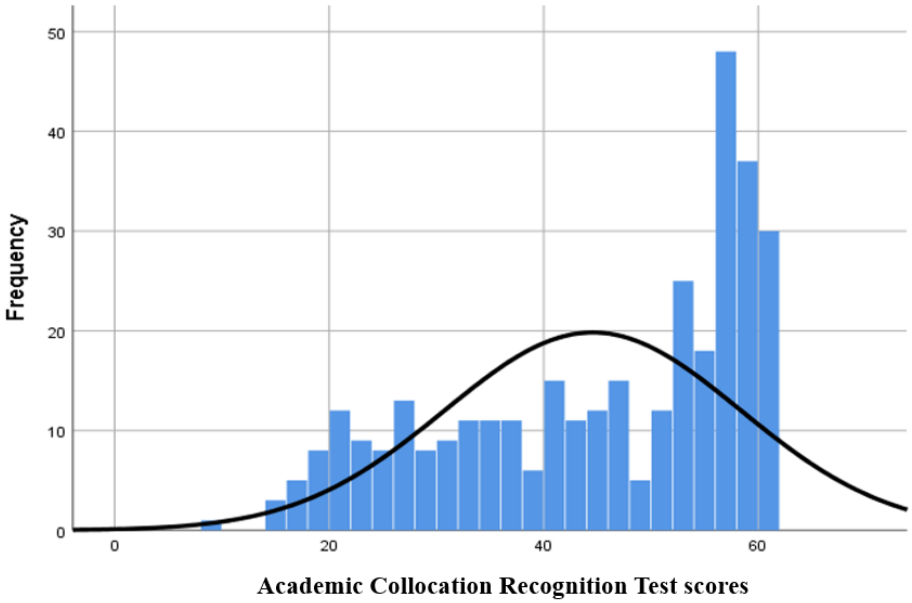
Summary Statistics of Test Scores (N = 343)

	Min	Max	Mean	Median	Standard deviation	Shapiro-Wilk *p* value
AC Recognition Test	9.00	60.00	44.57	48.00	13.80	<.001
AC Recall Test	0.00	58.00	26.03	25.00	14.82	<.001
VST	38.00	135.00	95.91	98.00	23.58	<.001

The score distributions of the three tests are visualised through the histograms in Figures 5.1, 5.2 and 5.3. For the ACTs, the distributions of the scores were different between the AC Recognition Test and the AC Recall Test. The histogram of the AC Recognition Test scores (Figure 5.1) was negatively skewed, suggesting that the majority of test-takers scored above average. In fact, approximately 80% of test-takers scored higher than 30 out of 60 on the AC Recognition Test, and almost 10% reached the maximum score of 60.

Figure 5.1

Score Distribution of the AC Recognition Test (N = 343)



By contrast, the test score set of the AC Recall Test (Figure 5.2) had better normality in distribution despite not being a perfect bell-curved shape. Its histogram was slightly skewed to the right, indicating that there were more lower scores than the mean score in the AC Recall Test. Nearly 40% of the test-takers scored above average and none of the test-takers achieved the maximum score for this test.

Figure 5.2

Score Distribution of the AC Recall Test (N = 343)

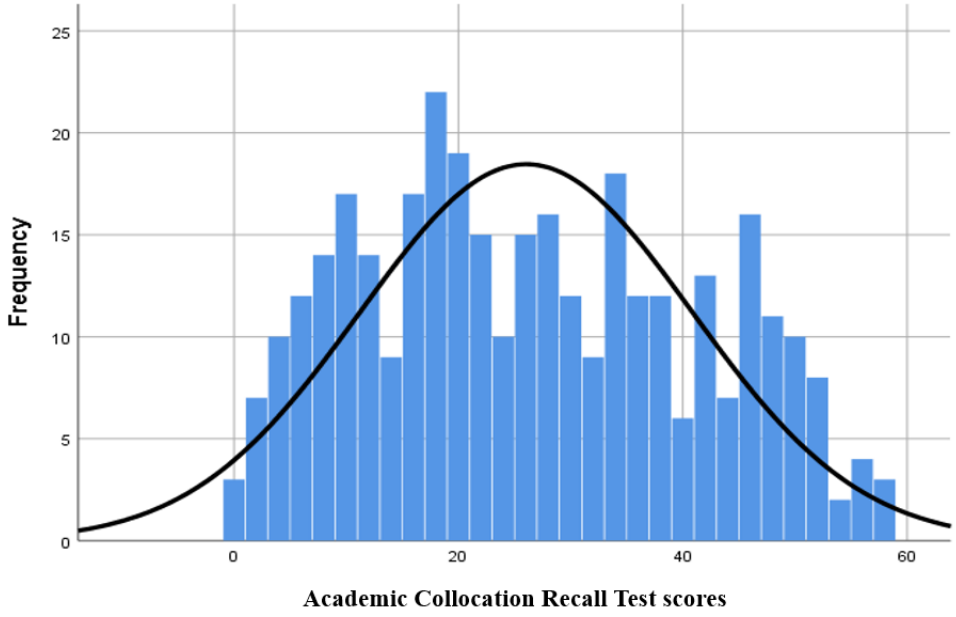
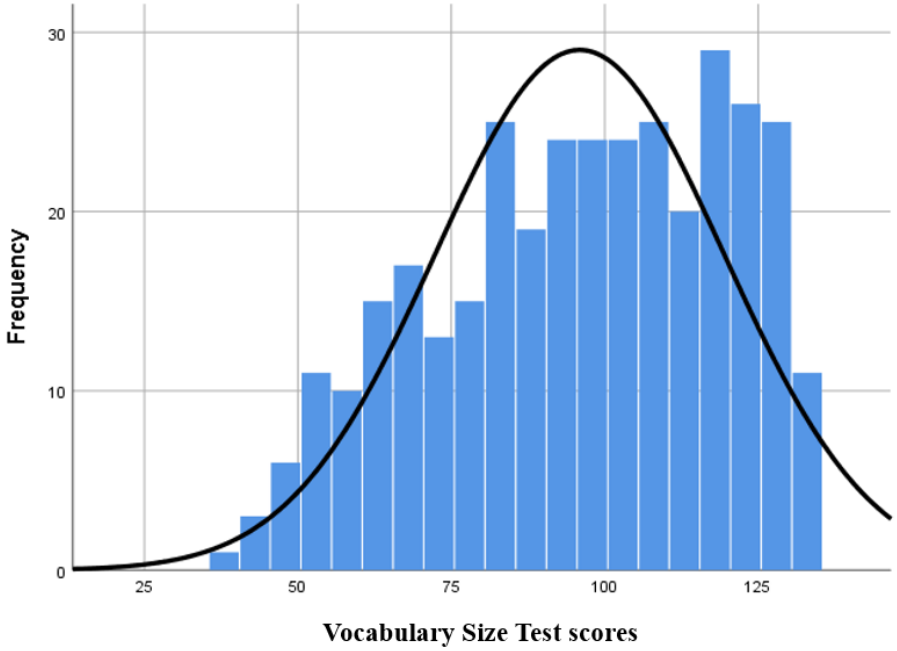


Figure 5.3 shows that there was a negative skewness of the score distribution for the VST scores, which suggests that a greater number of scores were higher than the mean score. Almost 50% of test-takers scored higher than 100 out of 140. Less than 3% of test-takers scored below 50. In other words, a vast majority of test-takers had a vocabulary size of larger than 5,000 word families and about half of the test-takers had a vocabulary size of larger than 10,000 word families.

Figure 5.3

Score Distribution of the Vocabulary Size Test (N = 343)



Overall, this analysis showed that all three score sets of the AC Recognition Test, the AC Recall Test and the VST covered a wide range of scores with a large amount of variation in the group of participants being tested.

5.2 Rasch model analysis of the ACTs

Rasch analysis using Winsteps software (Linacre, 2019) provided statistical tools to investigate the score sets and determine whether items on the ACTs work well together to measure the intended construct. If the test items conform to the model predicted by Rasch, a possible conclusion might be that the tests are measuring the single construct of academic collocations and there is no significant construct-irrelevant variance. As explained in Chapter 3 (Section 3.2.4.1), the following properties of the AC Recognition Test and the AC Recall Test were examined: item measure, item fit, point-measure correlations, unidimensionality and local independence. Each is reported in turn below: first for the AC Recognition Test, and then the AC Recall Test.

5.2.1 Rasch model analysis of the AC Recognition Test

5.2.1.1 Item measure of the AC Recognition Test

Rasch item measure helps to investigate the item difficulty (see Chapter 3, Section 3.2.4.1). As shown in Table 5.2, the item measures of the AC Recognition Test were from -2.20 logits to 1.29 logits, while the person measures varied from -1.96 logits to 5.56 logits. The item measures have a narrower range than the person measures, 3.49 logits and 7.52 logits, respectively (Table 5.2, last column). The mean of item measures is always set at 0.00 logit in Winsteps software by default, and on average the person ability (i.e., test-takers' knowledge of academic collocations) at 1.83 logits was higher than the item difficulty in the AC Recognition Test (Table 5.2, column 4). The results indicated that generally, the AC Recognition Test was easy for this group of test-takers.

Table 5.2

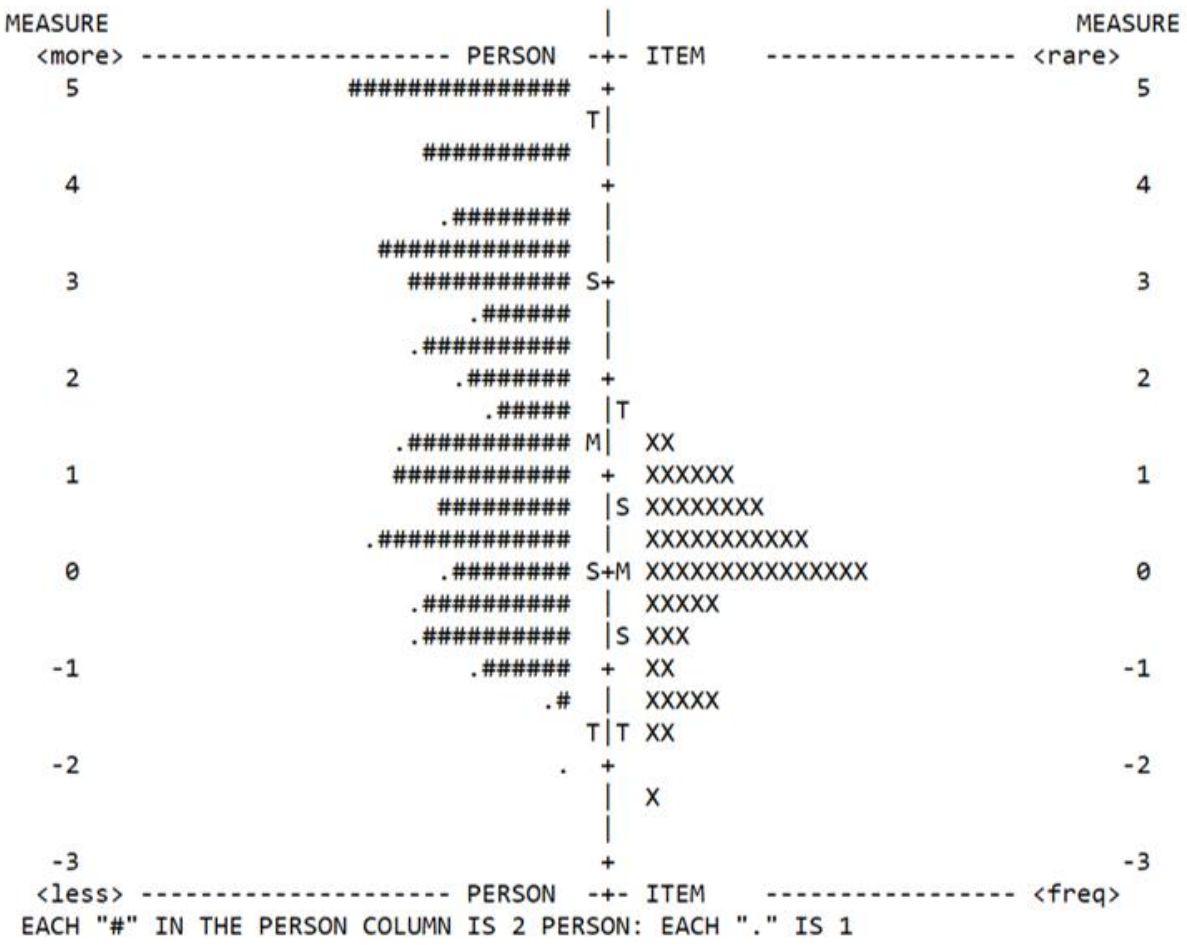
Rasch Measure of the AC Recognition Test in Logits

	Min	Max	Mean	SD	Range
Item measure	-2.20	1.29	0.00	0.81	3.49
Person measure	-1.96	5.56	1.83	1.88	7.52

The Wright Map in Figure 5.4 provides a visualisation of the relationship between test-takers' knowledge and item measures on the same scale of measurement. In this figure, the person column on the left refers to test-takers, and the item column on the right indicates the AC Recognition Test items. Persons with better knowledge of academic collocations (i.e., higher ability) and items that are more difficult are at the top of the map. Lower-performing persons and easier items are at the bottom of the map. As can be seen from Figure 5.4, the mean of person distribution lies above the mean of item distribution on the measurement scale. The majority of the AC Recognition Test items are clustered around a narrow band of person abilities, and the abilities of more than half of the test-takers were well above what the test could capture. In other words, the majority of the test-takers found the AC Recognition Test very easy.

Figure 5.4

Wright Map of the AC Recognition Test (343 Persons, 60 Items)



Note. M = mean of person or item distribution; S/T = One/ Two standard deviation(s) from the person or item mean.

5.2.1.2 Item fit of the AC Recognition Test

To gain insight into the test items, the fit statistics (infit MNSQ and infit ZSTD) of 60 AC Recognition Test items were examined. All the items whose MNSQ values are outside the range between 0.7 and 1.3 and whose ZSTD values are not from -2.0 to 2.0 were identified as being misfits. Because misfit may indicate problematic items, further investigation of these items is warranted. Detailed fit statistics of all the AC Recognition Test items can be found in Appendix E. A total of 11

misfit items were detected in the test as presented in Table 5.3 (misfit values are in bold). Among these, five items (11, 31, 38, 39 and 47) were overfit (ZSTD < -2.0) and six items (10, 18, 19, 25, 37 and 53) were underfit (MNSQ > 1.3 and ZSTD > 2.0).

Table 5.3

Misfit Items in the AC Recognition Test

Item	Infit MNSQ	Infit ZSTD
10	**1.31**	**4.17**
11	0.80	**-3.12**
18	1.27	**3.70**
19	1.24	**3.35**
25	1.23	**3.11**
31	0.80	**-2.86**
37	1.26	**3.42**
38	0.76	**-3.82**
39	0.81	**-2.90**
47	0.84	**-2.34**
53	**1.53**	**6.72**

Further investigation into item distraction efficiency of the AC Recognition Test items with unsatisfactory fit statistics was conducted to check if distractors could be a source of the misfit. Table 5.4 gives detailed information about options for each misfit test item. Taking item 11 in Table 5.4 as an example, it can be seen that options A, B and C as distractors (scored 0) were chosen by 30, 26, and 34 people, respectively. Option D as the correct answer (scored 1) was chosen by 252 people. One person did not give an answer for this item (i.e., blank). The “Point-measure correlation” column in Table 5.4 is the correlation between the response and person knowledge. The correlation for the correct option should be positive, and the correlation for the distractors should be negative (Linacre, 2019). The “Ability mean” column refers to the average measure of persons who responded with a given option. It is expected that higher ability people selected correct options and lower ability people selected distractors. As a result, distractors need considering to be replaced if:

- their point-measure correlation was positive;
- the mean ability of people choosing the distractor was higher than that of the correct answer;
- they were not selected by any test-takers, indicating that the wrong option was too obvious.

It can be seen from the last column in Table 5.4 that all the distractors had a negative point-measure correlation, indicating that the response of test-takers followed the expected pattern. All the distractors were selected by people with lower abilities than those who selected the correct answers (see column 6, Table 5.4). Moreover, all the distractors attracted the test-takers, suggesting their quality. If a distractor looked clearly wrong or irrelevant, no test-taker would consider it. As a result, no distractor needs replacing. On the other hand, there were seven items (11, 25, 37, 38, 39, 47, and 53) for which some test-takers did not answer (see column 2, Table 5.4). Although the count for blank answers was no more than 1% of the total count for each item, these answers might still have an impact on the item fit.

Table 5.4

Distractor Analysis of Misfit AC Recognition Test Items

Items	Options	Score value	Data count	%	Ability mean	Point-measure correlation
10	C	0	26	8	0.23	-0.24
	B	0	66	19	0.83	-0.26
	D	0	37	11	0.9	-0.17
	A	1	214	62	2.5	0.46
11	A	0	30	9	-0.12	-0.32
	B	0	26	8	-0.08	-0.29
	C	0	34	10	0.01	-0.32
	Blank	0	1	0	0.81	-0.03
	D	1	252	73	2.52	0.6
18	C	0	13	4	-0.27	-0.22
	D	0	24	7	0.37	-0.21
	B	0	81	24	0.86	-0.29
	A	1	225	66	2.46	0.46
19	B	0	34	10	0.33	-0.26
	C	0	47	14	0.63	-0.26
	A	0	40	12	0.91	-0.18
	D	1	222	65	2.49	0.47
25	A	0	28	8	0.6	-0.2
	C	0	53	15	0.71	-0.26
	D	0	61	18	0.75	-0.27
	Blank	0	1	0	0.81	-0.03
	B	1	200	58	2.64	0.51

Items	Options	Score value	Data count	%	Ability mean	Point-measure correlation
31	C	0	13	4	-0.38	-0.23
	B	0	20	6	-0.22	-0.27
	A	0	37	11	-0.16	-0.37
	D	1	273	80	2.36	0.55
37	Blank	0	2	1	-0.29	-0.09
	C	0	23	7	-0.18	-0.29
	D	0	13	4	0.68	-0.12
	A	0	110	32	0.99	-0.31
	B	1	195	57	2.65	0.5
38	Blank	0	1	0	-0.6	-0.07
	B	0	16	5	-0.39	-0.26
	A	0	37	11	-0.19	-0.37
	C	0	38	11	0.17	-0.31
	D	1	251	73	2.54	0.62
39	Blank	0	1	0	-0.6	-0.07
	B	0	16	5	-0.39	-0.26
	A	0	47	14	0.15	-0.36
	C	0	44	13	0.23	-0.33
	D	1	235	69	2.64	0.63
47	Blank	0	1	0	-0.6	-0.07
	C	0	19	6	-0.25	-0.27
	B	0	21	6	-0.25	-0.28
	A	0	39	11	0.12	-0.33
	D	1	263	77	2.42	0.56
53	B	0	41	12	0.76	-0.21
	A	0	36	10	0.85	-0.18
	D	0	36	10	1.2	-0.12
	Blank	0	4	1	1.62	-0.01
	C	1	226	66	2.29	0.34

In sum, there were 11 misfit items in the AC Recognition Test identified by Rasch analysis, but the source of any misfit could not be detected. The distractor analysis did not point out any problems except for some blank answers that may have affected the fit statistics. The context sentences which had been carefully checked and piloted should not be an issue either. Among 11 misfit items, five overfit items (11, 31, 38, 39, and 47) should not cause concern. Overfit items do not degrade the quality of the instruments (Linacre, 2019) and could be the best items of the test as the other items are not as discriminating as these (Wu et al., 2016), especially when no source of misfit could be identified. Because this is not a high-stakes test, 10% of underfit items (10, 18, 19, 25, 37, and 53) should not cause a serious threat to the overall quality of the test for its diagnostic purpose.

5.2.1.3 Point-measure correlations of the AC Recognition Test items

Positive point-measure correlations (r >.20) are expected so that responses to test items align with person abilities (see Chapter 3, Section 3.2.4.1). The last column of Appendix E provides the correlation for each test item of the AC Recognition Test. Overall, the correlations were all positive (from .30 to .63), suggesting that the test items were aligned in the same direction to measure the intended construct.

5.2.1.4 Unidimensionality of the AC Recognition Test

So far, the fit statistics and the point measure correlations have provided some indication of the unidimensionality of the AC Recognition Test. The test items have shown to be relatively fit to the Rasch model and they work together to measure the construct of academic collocation knowledge. Principal component analysis of residuals (see Chapter 3, Section 3.2.4.1) was further conducted to address the dimensionality issue more directly. In the AC Recognition Test, the first contrast had an eigenvalue of 2.4 and accounted for the small variance of 2.8% in the data. According to Linacre (2019), the contrast might not be an additional dimension but might be simply a result of a random effect in the data. This seems to be the case when the analysis of the local independence below did not point out significant standardised residual correlations between any item pair.

5.2.1.5 Local independence of the AC Recognition Test items

The local independence of test items is identified by Rasch residual correlations via Q3 coefficients (see Chapter 3, Section 3.2.4.1). The largest Q3 coefficients are presented in Table 5.5. As shown in the table, the correlation values of all the item pairs fell well within the suggested range between -0.3 and 0.3 (Fan & Bond, 2019), except for the first pair of items 2 and 38 (r = .32) whose value was slightly higher. Looking more closely at these two items (see Figure 5.5), the wordings are very different, and test-takers could scarcely form any association between them. In other words, the answer to one item would not affect how the other item was answered. Therefore, it can be said that items 2 and 38 are sufficiently independent of each other. The findings suggest that the requirement of local independence is held for the AC Recognition Test.

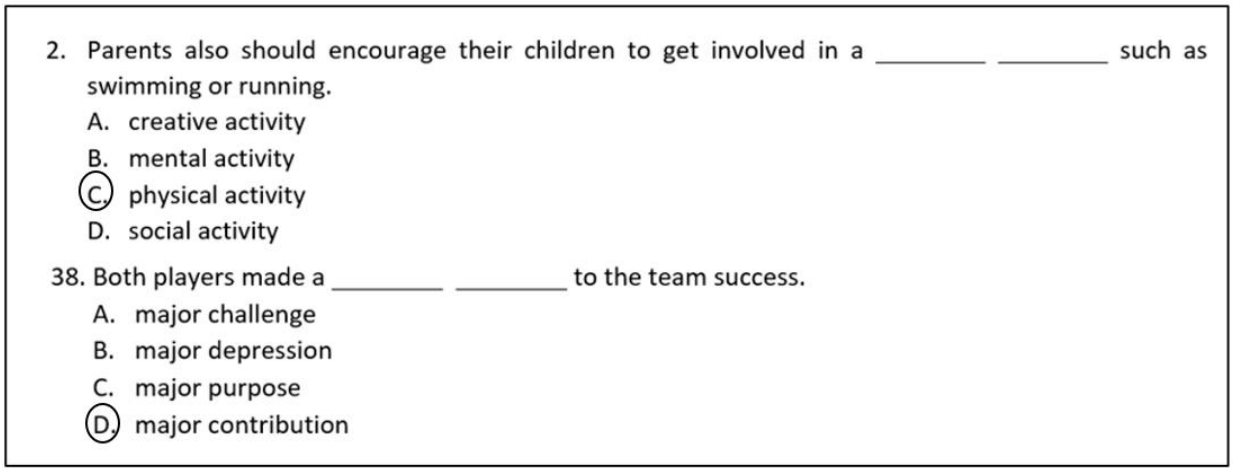
Table 5.5

Largest Standardised Residual Correlations of the AC Recognition Test Items

Item	Item	Q3 coefficient
2	38	0.32
31	38	0.29
12	26	0.24
5	46	0.22
13	48	0.21
4	7	0.20
5	22	0.19
18	27	0.18
21	24	0.18
3	51	0.18
24	45	0.17
7	11	0.17
47	48	0.17
40	45	-0.19
8	46	-0.18
37	38	-0.17
27	54	-0.17
8	33	-0.17
4	54	-0.16
7	54	-0.16

Figure 5.5

Items 2 and 38 of the AC Recognition Test



The Rasch analysis has pointed out that the AC Recognition Test items acceptably fit the Rasch model. Although a few items were identified as being misfits, they would not pose a serious threat to the validity of the test results. Following the same analysis procedure, the next section presents the Rasch analysis results of the AC Recall Test.

5.2.2 Rasch model analysis of the AC Recall Test

5.2.2.1 Item measure of the AC Recall Test

The items of the AC Recall Test showed a wide range of item measures from -4.27 logits to 4.72 logits (see Table 5.6), which indicates that the items in the AC Recall Test are very different in terms of item difficulty. The mean item measure was set at 0.00 logit by default and a lower mean at -0.49 logit of person measure suggested that the AC Recall Test was difficult for this group of candidates. The Wright Map in Figure 5.6 gives a fuller picture of the alignment of the item difficulty with the person ability. Figure 5.6 clearly shows that the distribution of items on the right relatively matched the distribution of persons on the left, although a few people seemed to have lower recall knowledge of academic collocations than what the test could capture.

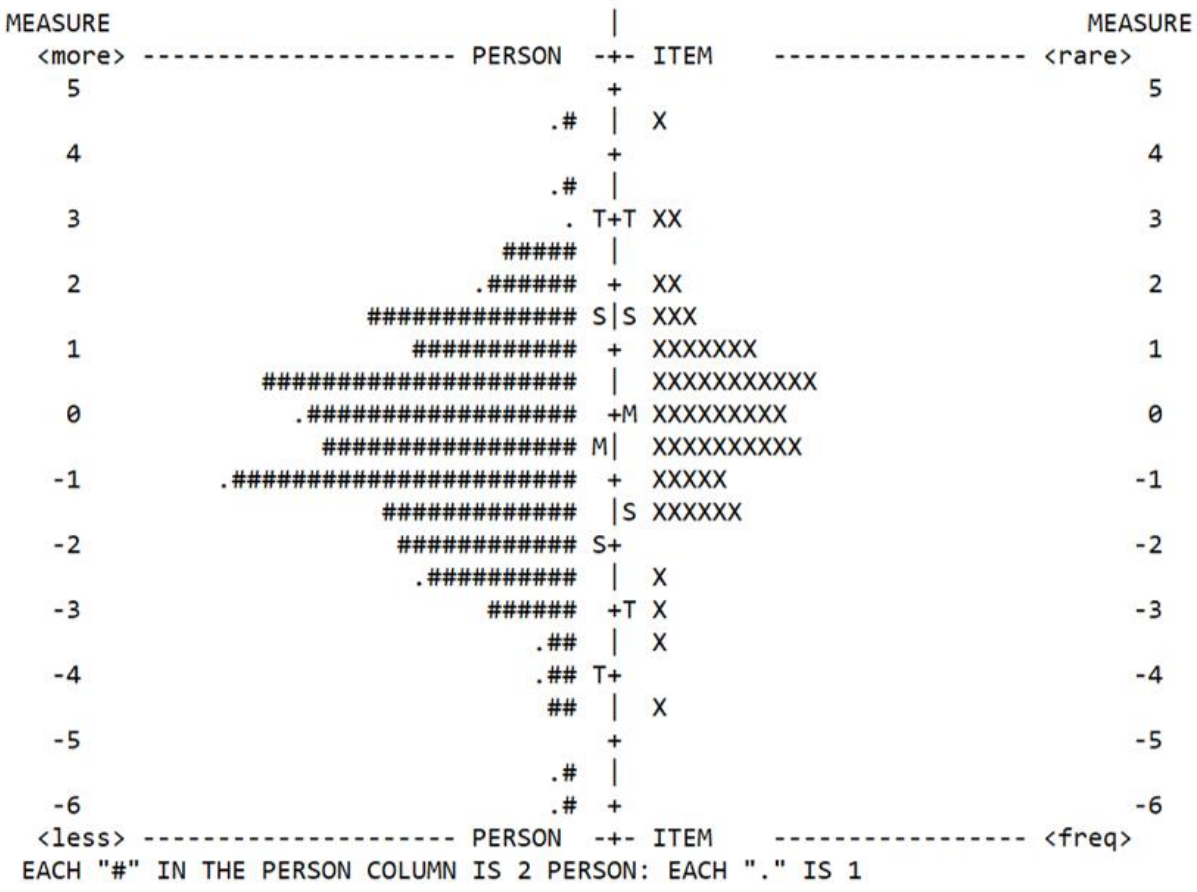
Table 5.6

Rasch Measure of the AC Recall Test in Logits

	Min	Max	Mean	SD	Range
Item measure	-4.72	4.27	0.00	1.47	8.99
Person measure	-6.64	4.33	-0.49	1.8	10.97

Figure 5.6

Wright Map of the AC Recall Test (343 Persons, 60 Items)



Note. M = mean of person or item distribution; S/T = One/ Two standard deviation(s) from the person or item mean.

5.2.2.2 Item fit of the AC Recall Test

As explained in Chapter 3 (Section 3.2.4.1), items in the AC Recall Test were regarded as misfits if their MNSQ values did not fall between 0.5 and 1.5, or their ZSTD values were not from -2.0 to 2.0. Detailed fit statistics of all the AC Recall Test items are presented in Appendix F. Table 5.7 shows three items of the AC Recall Test which were identified as misfits: one overfit item (14) and two underfit items (6 and 7). The overfit item could be the best item of the test because it is highly discriminating (Wu et al., 2016). According to Smith Jr (2005), having less than 5% of underfit items

would not affect the overall quality of the measurement. As a result, the AC Recall Test items fit the Rasch model well.

Table 5.7

Misfit Items in the AC Recall Test

Item	Infit MNSQ	Infit ZSTD
6	1.28	**3.59**
7	1.18	**2.60**
14	0.76	**-2.68**

5.2.2.3 Point-measure correlations of the AC Recall Test items

The point measure correlations were examined to check whether the items of the AC Recall Test aligned to the same direction. Linacre (2019) suggests that negative or low correlations ($r < .20$) signal problematic items. Appendix F shows the correlation value for each AC Recall Test item (see the last column). All the positive point measure correlations (from .28 to .67) indicated that items in the AC Recall Test work well together to measure the intended construct.

5.2.2.4 Unidimensionality of the AC Recall Test

The fit statistics and point-measure correlations have provided some evidence of unidimensionality of the AC Recall Test items which were confirmed by the principal component analysis of residuals. In this analysis, a first contrast with an eigenvalue above 2 was considered substantive enough to represent a component (Linacre, 2019). The first contrast of the AC Recall Test had an eigenvalue of 2.5 and accounted for the small variance of 2.4% in the data, suggesting that there might be an additional dimension being measured by the instrument. In reality, a complete unidimensionality can scarcely be achieved, and the fit statistics help to identify where the problems can be found. The above analysis of misfit items did not point out any particular concerns. Therefore, it is likely that the AC Recall Test has no substantial issues of unidimensionality. The following local independence analysis will also confirm the unidimensionality of the test.

5.2.2.5 Local independence of the AC Recall Test items

As mentioned in Chapter 3 (Section 3.2.4.1), local independence (i.e., test items are independent of each other) can be investigated via values of Q3 coefficients. Table 5.8 shows that the Q3 coefficients of the item pairs in the AC Recall Test were very low (from -0.17 to 0.30). This finding suggested that no pair of items in the test has a strong relationship to constitute a secondary construct of the measure. All the test items are independent so that the answer of one item does not affect how the other items are answered.

Table 5.8

Largest Standardised Residual Correlations of the AC Recall Test Items

Item	Item	Q3 coefficient
44	46	0.30
49	56	0.23
11	22	0.22
10	28	0.22
46	48	0.21
14	27	0.19
41	43	0.18
16	21	-0.20
39	44	-0.19
20	38	-0.19
3	55	-0.19
4	33	-0.19
9	41	-0.18
31	53	-0.18
8	39	-0.18
15	38	-0.18
24	53	-0.18
20	39	-0.17
6	47	-0.17
51	55	-0.17

In sum, the AC Recall Test conformed to the Rasch model with items that are locally independent and aligned to measure the unidimensional knowledge of academic collocations. Taken as a whole, the findings presented in Section 5.2 partly reveal the answer to RQ1, which confirms that items of the ACTs showed proper fit to the Rasch model. In other words, the statistical characteristics of the

test items are appropriate to measure the intended construct. The next section expands on this section by providing qualitative evidence to add to the assessment of the test characteristics.

5.3 Test-takers' opinions about the ACTs

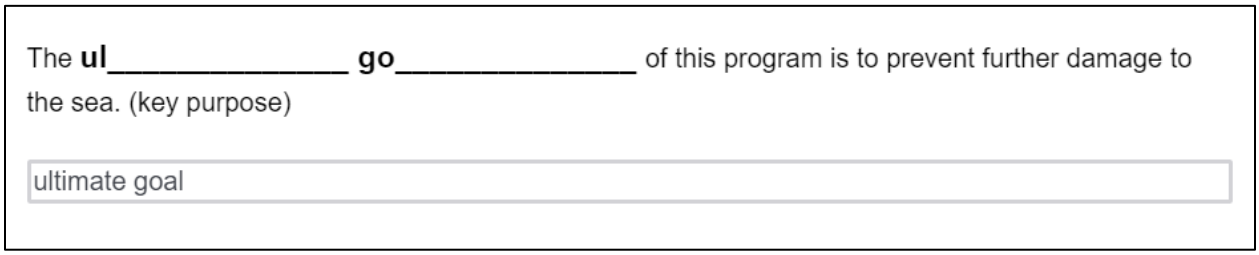
Test-takers' comments on the ACTs, which were elicited with questions in the post-test interview (see questions 1 to 3, Section B, Appendix B), helped to evaluate whether the test characteristics worked as intended. All the respondents thought the AC Recall Test was more difficult than the AC Recognition Test, as expected. This finding matched the statistical analysis in Section 5.2 which showed that generally, the AC Recognition Test was easy, while the AC Recall Test was difficult for the participants in this study. Although the respondents were not directly asked about the test format, they tended to refer to the format when being asked to give comments on the tests or talk about the test difficulty. The opinions about the test format seem to have greater significance than the test difficulty; therefore, the format is the focus of this section.

5.3.1 Opinions on the AC Recall Test format

Respondents commented on three elements of the AC Recall Test format, namely, the sentence prompts, the two-initial-letter hints, and the meanings of the collocations. Figure 5.7 (also used as Figure 4.6 in Chapter 4) is presented below as an example. Test-takers were asked to provide a two-word collocation to fit the given context. The two initial letters were consistently provided for each word of the collocation pairs. The meanings of the elicited collocations were put in brackets at the end of the context sentences.

Figure 5.7

Example of an AC Recall Test Item



Firstly, a majority of the interviewees (37/44) provided positive feedback on the sentence prompts. They reported that the sentences were "very clear", "easy to understand", "straightforward", "understandable" or "reasonable". This feedback means that the context sentences acted as intended. Only seven respondents revealed uncertainty with some sentence prompts. The reason given by Ngoc VN, a second-year undergraduate student, was that:

> *I found the context sentences more difficult to understand in the first test* [AC Recall Test] *with the blanks because the meaning was incomplete* (Excerpt 5.1).

The second element of the AC Recall Test format – the two-initial-letter hint – was perceived differently by the participants. Only 11 respondents commented on this feature of the test. Among them, three acknowledged the usefulness of the initial letters. For example, Dieu VN – a second-year student – reported that:

> *For me, with a phrase that I know, the initial letters helped me to recall it more quickly, but with a phrase I didn't know, even with the initial letters, I still didn't know.* (Excerpt 5.2)

Another four respondents stated that providing only two initial letters for each word was not enough for them to produce the needed collocations. A third-year student, Kim VN, said that:

> *If there had been more initial letters provided, the first test* [AC Recall Test] *could have been easier.* (Excerpt 5.3)

On the other hand, the other four respondents thought that the initial letter hints did not help much. Nhi VN – a third-year student – explained that:

> *Without the two initial letters, various answers could have been accepted, and the difficulty level would have been decreased. With the initial letters, test-takers might not remember exactly the phrases needed in that case.* (Excerpt 5.4)

The participants' opinions were valuable for pointing out both pros and cons of the test format with the provided initial letters. While these letters helped to elicit the intended collocations, they also

restricted the possible answers, which might cause difficulty for test-takers when they only knew equivalent phrases but not the elicited collocations.

Thirdly, the meanings in brackets of the academic collocations seem to be an important feature of the AC Recall Test. The respondents referred to them with favourable attitudes, as can be seen in the following representative comments:

> *The meanings in the brackets were very close to the needed phrases. There were cases I couldn't guess the answers when reading the context sentences, but the meanings in the brackets helped me know the answers.* (Dieu VN, Excerpt 5.5)

> *The meanings in the brackets were useful because the meanings were the same as the needed phrases so I could infer the answers.* (Thuong VN, Excerpt 5.6)

Notably, one interviewee commented on the position of the collocation meanings. Ho NZ – a PhD student in Education – said:

> *At first, I didn't notice the meanings in the brackets. After a while, I realised that the words in brackets gave hints for the needed phrases. When doing the tests, I usually didn't read the whole sentences. That's why I missed some words in brackets.* (Excerpt 5.7)

Although the meaning hint was clearly mentioned in the test instruction, there was still a chance that test-takers did not read the instruction carefully.

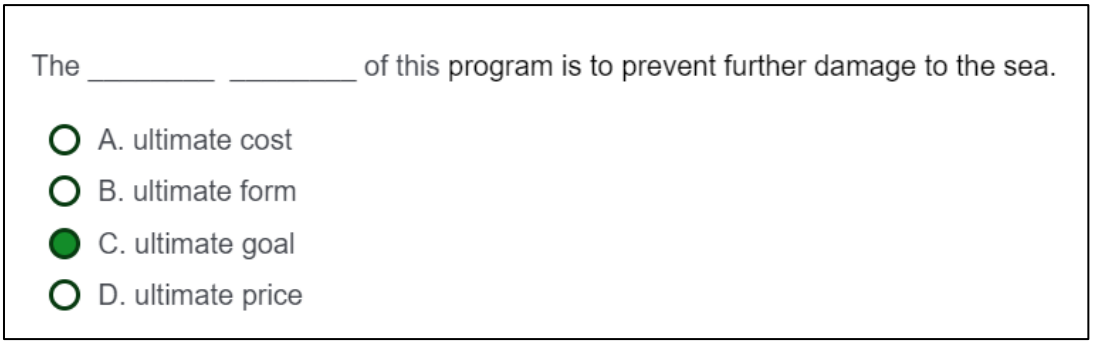
Overall, the features of the AC Recall Test format seemed to work well. The context sentences written in plain language were perceived positively by most of the interviewees as straightforward and comprehensible. This is important as these meaningful context sentences were intended to help test-takers to infer the needed collocations to fill in. Although the initial letter hints might cause difficulty to some test-takers, these hints still fulfilled their task to limit possible correct answers and elicit the intended academic collocations. Finally, the provided meanings of collocations were also proved to be useful. The following section turns to test-takers' comments on the format of the AC Recognition Test.

5.3.2 Opinions on the AC Recognition Test format

Multiple-choice is a familiar test format for learners from various first language backgrounds; hence, it is not surprising that the participants tended to favour the AC Recognition Test. Figure 5.8 (also used as Figure 4.5 in Chapter 4) is presented to illustrate the AC Recognition Test format. Test-takers were asked to select a suitable academic collocation to fill in the two-word blank.

Figure 5.8

Example of an AC Recognition Test Item



The AC Recognition Test format received no criticism from the interviewees. Their comments all reflected the ease of the format in giving the answers, such as the following:

> *We had the options provided. We didn't have to remember the spelling, so it was easier.* (Hong NZ, Excerpt 5.8)

> *When I saw the choices laid down side by side, it was easier for me to decide which one was wrong and which one was probably true.* (Linka NZ, Excerpt 5.9)

> *The third test* [AC Recognition Test] *repeated the same context sentences but it was easier to guess the meaning.* (Ngoc VN compared the format of the AC Recognition Test against the AC Recall Test, Excerpt 5.10)

> *I could see the options in a familiar context, so I could choose easily.* (Dung NZ, Excerpt 5.11)

From the above quotations, it is reasonable to conclude that the multiple-choice is a suitable format for the purpose of measuring recognition knowledge of academic collocations. Test-takers easily recognised the collocations that they knew among the given options (Excerpts 5.8 and 5.9). The same context sentences being used in both the AC Recall Test and the AC Recognition Test also supported test-takers, as mentioned by Ngoc and Dung (Excerpts 5.10 and 5.11). This feature of the test worked as intended to reduce the reading burden for test-takers when they had to take the two collocation tests at the same time.

An important feature of the AC Recognition Test was acknowledged by Trang NZ who showed a preference to the current format to test knowledge of academic collocations:

> *I like the way that the phrases go together in your tests rather than being separated. For example, if I gave a wrong answer in a matching test, I would easily remember that incorrect combination. Your tests avoid that issue because the words go together.* (Excerpt 5.12)

If we look back at the test development process in Chapter 4 (Section 4.4), the point made by Trang NZ (Excerpt 5.12) had been taken into account. The collocations were intentionally presented as whole units to raise learners' awareness about this group of multiword units. Moreover, even the distractors in the AC Recognition Test are real academic collocations. The test, therefore, has no side effect on test-takers' memory trace.

In sum, the respondents' opinions about the AC Recognition Test format were positive. Participants were comfortable with the familiar multiple-choice format. The repetition of context sentences from the AC Recall Test to the AC Recognition Test helped to ease the reading burden. The representation of the collocations as whole units was also recognised as an advantage of the test format. The following section delves deeper into the effectiveness of the test format by examining the participants' test-taking strategies.

5.4 Test-takers' reflections on the test-taking strategies

The test-taking strategies presented in this section reveal the cognitive processes that test-takers went through when answering test items. These shed light on whether the ACTs were able to elicit the intended knowledge, and whether the observed test results appropriately reflected participants'

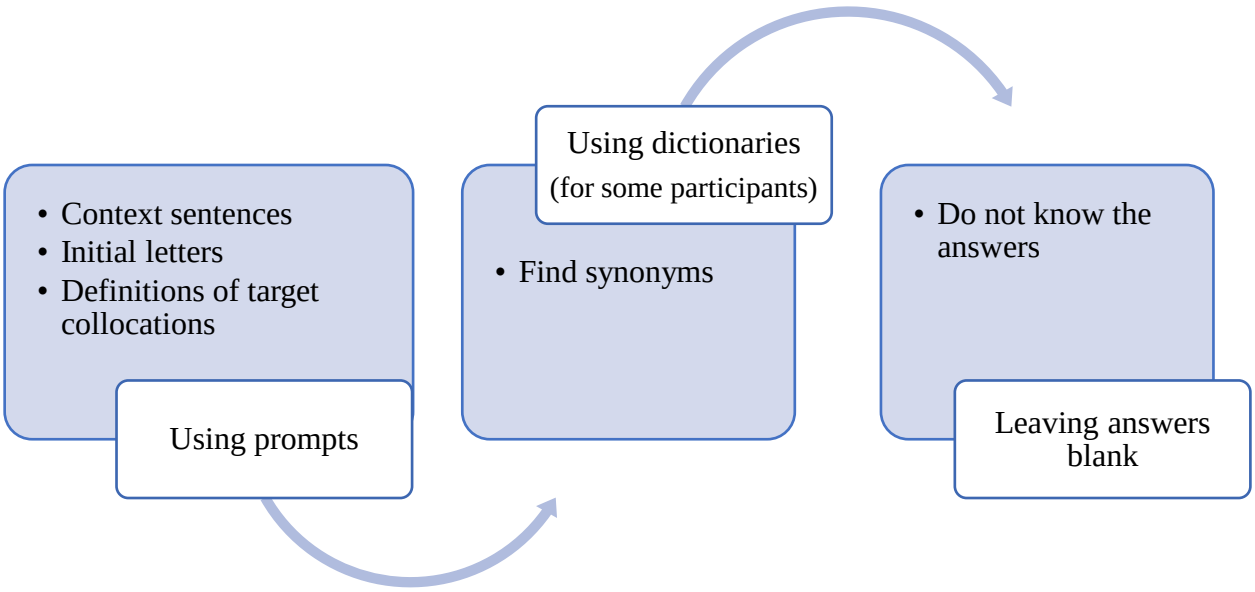
knowledge of academic collocations. The strategies reported by test-takers also contain some information about the online testing condition. These strategies were revealed by the interviewees through responses to direct questions (4-7 in Section B, Appendix B) and further explanation during the process of re-taking the ACTs verbally (Section C, Appendix B).

5.4.1 The AC Recall Test strategies

The respondents reported using three test-taking strategies for the AC Recall Test: using the prompts, using dictionaries and leaving answers blank. The first strategy was commonly used by all the test-takers. The second was employed by about a third of participants only when the prompts did not help them to find the answers. Leaving blanks was the last option when test-takers could not think of any response to fill in. These strategies are sequenced as illustrated in Figure 5.9 and will be discussed in turn.

Figure 5.9

Strategies Reported by Interviewees for the AC Recall Test



5.4.1.1 Using prompts

The prompts of the AC Recall Test, as illustrated in the example in Figure 5.7, include the context sentences, the two initial letters, and the definitions in brackets of the target academic collocations. Test-takers relied on those prompts to answer test items. Interestingly, participants did not make use of the prompts in the same way, as illustrated in Excerpts 5.13 to 5.15. Test-takers chose different prompts to focus on first. Tran NZ (Excerpt 5.13) started with reading the context sentences, while Dung NZ (Excerpt 5.14) looked at the definitions of collocations and Naomi NZ (Excerpt 5.15) relied on the initial letter hints first. It is important to note that all the hints were employed by all the participants in some ways.

> *I read the whole sentence first, but not the initial letter hint. I tried to find the phrase on my own first. If it matched the hint, good. Otherwise, I had to find another one.* (Tran NZ, Excerpt 5.13)

> *To do this test* [AC Recall Test]*, I read the definition in the brackets first, then I guessed the part of speech of the needed phrases; for example, the first word was an adjective, and the second word was a noun. Then I read the whole context sentence to guess the phrase. If I didn't know the answer, I would find the phrase that had the closest meaning.* (Dung NZ, Excerpt 5.14)

> *I would pronounce the two letters in my head, trying to see if they could bring anything about. And then when I could understand the context better, some words came in my head …* (Naomi NZ, Excerpt 5.15)

It is important that the test-takers drew on the content and format of the AC Recall Test rather than external resources (e.g., dictionaries) to answer the questions because this indicated that the test results accurately reflected learners' recall knowledge of academic collocations. Using the prompts, however, did not guarantee that test-takers were able to answer the test items. At that point, some of the test-takers turned to dictionaries.

5.4.1.2 Using dictionaries

Referring to any external sources when taking the ACTs would cause a threat to the validity of the test results. About a third of the respondents (16/44) reported using dictionaries to some extent to look up the answers for the AC Recall Test. No other sources were mentioned by the interviewees. To evaluate whether the use of dictionaries seriously affected the validity of the test results, the following three threads that arose from the data were reported in return:

- Frequency of checking the dictionaries
- How the dictionaries were used to look up the answers
- Effectiveness of using dictionaries in answering the test

First, the frequency of dictionary consultation varied across the respondents, but it was clear that no interviewee looked up every test item. Nhi VN and Thom VN reported that they relied on dictionaries "quite often" during the test. The other interviewees described the frequency of dictionary use as "not much", "for a couple of items", "just a little" or "only for one word". Participants seemed to be selective in using dictionaries. For example, Cam VN – a third-year undergraduate said:

> *I used the dictionary only for items that I could guess one word. I looked up the other word then.* (Excerpt 5.16)

Second, it is important to note that the respondents were not able to give many examples of items they had looked up in dictionaries while they were taking the tests. They were also unable to clearly describe how they went about consulting the dictionary for the collocations. Excerpt 5.17 is a representative comment of the way most participants reported that they had made use of the dictionaries. Nhi VN (Excerpt 5.18) was the only participant who could share a specific example of the test item found in a dictionary.

> *I used an online thesaurus for the first test [AC Recall Test]. I could guess the meaning, but I didn't know the exact phrase that matched the initial letters provided. It was quite complicated, so I needed to use the thesaurus to find synonyms to fill in*. (Dung NZ, Excerpt 5.17)

> *I searched for an English phrase. For example, for the item 'free movement', I searched the meaning in English – what word means 'free to move', and then the phrase appeared. Then I compared the two initials and if they were the same, I would use that phrase.* (Nhi VN, Excerpt 5.18)

Respondents reported that they tended to rely on the definitions of the academic collocations provided in the dictionaries and used thesaurus dictionaries to find equivalent phrases. "Thesaurus" and "synonyms" were the two keywords that were frequently mentioned in the interviews, suggesting that the way the test-takers searched for the needed academic collocations was similar to looking up single words.

Third, the interviewees were clear that consulting dictionaries was not a very effective strategy for finding answers for the AC Recall Test. The respondents expressed their disappointment when using dictionaries in these ways:

> *I looked up a couple of items in the first test* [AC Recall Test] *but it was in vain, so I did the rest on my own.* (Khanh VN, Excerpt 5.19)

> *It was extremely difficult. To look up in the dictionary, I needed to search in a logical way, and it took a really long time.* (Thuy VN, Excerpt 5.20)

> *Yes, I searched for a couple of items, but then I thought 'oh, this is gonna take forever' so I just did the test as to my best knowledge.* (Linka NZ, Excerpt 5.21)

> *I used the thesaurus to find synonyms, but I couldn't find anything. It wasn't effective.* (Morgan NZ, Excerpt 5.22)

It is not surprising when dictionaries seemed not to help much. Hong NZ who realised this issue with dictionaries commented that:

> *If I had used the dictionary, I would have needed to look up two words. Otherwise, looking up only one word was nonsense. So, I didn't use the dictionary.* (Excerpt 5.23)

The use of thesaurus dictionaries to search for synonyms, as the respondents shared in Excerpts 5.19 to 5.22, might only be effective for single words. With collocations, it can be very easy to find the meaning of a collocation with dictionaries given that its meaning is transparent and can be inferred from its components. For instance, learners might in turn look up the words “ultimate” and “goal” to know the meaning of the collocation “ultimate goal”. The other way round, starting with the meanings and finding the collocation, would be difficult. To the best of my knowledge, currently, there is no English dictionary or bilingual dictionary that supports learners to search for collocations from a provided meaning.

In brief, although the use of dictionaries in this testing context acted against the Evaluation inference, the effect of the dictionary use on the test results might not be particularly impactful. This was because the dictionaries were not used often, and it was difficult to find the answers even with dictionaries.

5.4.1.3 Leaving answers blank

As shown in Figure 5.7, leaving answers blank was the last option when test-takers could not rely on the prompts and dictionaries. Blank answers accounted for approximately 20% of the total responses in the AC Recall Test (i.e., 3,937/20,580). Among 44 respondents, two persons reported skipping one or two items by accident (Tien VN and Thu VN). A total of 12 people did not leave any blank answers. The other 30 participants confirmed that blank responses meant that they did not know the answers. This is illustrated by the comment of Kaylee NZ:

> *For the first test* [AC Recall Test], *I did leave some items blank because I was just not sure, or I didn’t know about them at all.* (Excerpt 5.24)

How might blanks affect scoring the test? If test-takers skipped answering without reading a test item, a judgement could not be made on whether they had the knowledge measured by that test item or not. If test-takers knew the answers but left blanks and received zero points, the test scores would underestimate their knowledge. The interview data shed some light on this question. The data showed that the majority of blank answers were an indication from the test-takers that they did not know the academic collocations that were being tested in particular items. This finding supported the scoring rubric in which blank answers were treated as incorrect answers and received zero points.

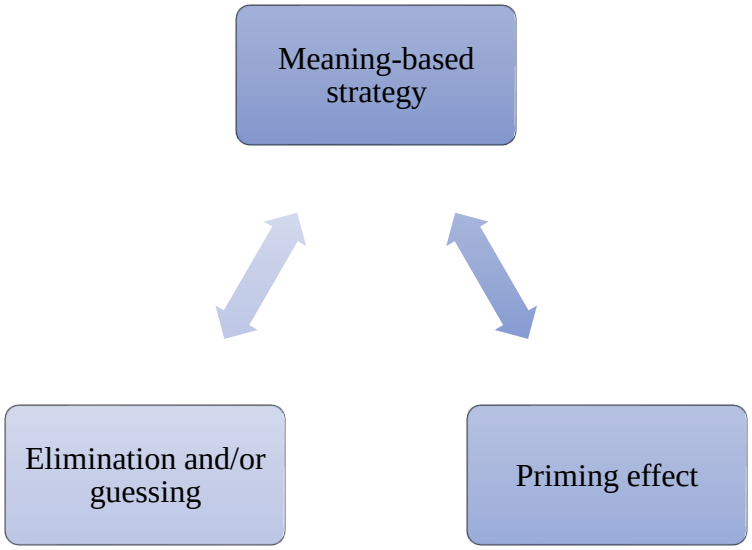
Taken together, the respondents' reflections on the strategies to take the AC Recall Test provided evidence for and against the Evaluation inference. On the one hand, this inference was supported because using the prompts was the main strategy of all the participants to answer the test. This suggested that the test allowed test-takers to demonstrate the target knowledge and the test-taking process was appropriate. On the other hand, the fact that some participants used dictionaries to find answers for test items or test-takers skipped test items without reading them casts doubt on the validity of the test results. That said, little evidence was found on the effectiveness of using dictionaries in providing correct responses for the AC Recall Test and very few blanks were left without any effort in answering the test items. Looking into the strategies employed by test-takers for the AC Recognition Test in the following section provides more evidence for the overall assessment of the Evaluation inference.

5.4.2 The AC Recognition Test strategies

For the AC Recognition Test, the respondents reported using three strategies: meaning-based strategies, priming effect, and elimination and/or guessing. The first strategy was employed by all the participants, while the latter two were utilised only when they could not infer answers from sentence prompts and options given. In general, the main approach was meaning-based strategy, and it provided the foundation for the other two strategies. Figure 5.10 illustrates the relationship between the strategies for the AC Recognition Test. It should be noted that only 6/20,580 answers were left blank (0.03%), which showed a non-significant trend in giving responses for the AC Recognition Test. Blank responses were therefore not included in the analysis for this test.

Figure 5.10

Strategies Reported by Interviewees for the AC Recognition Test



5.4.2.1 Meaning-based strategies

The most common explanation for participants' responses on the AC Recognition Test was that they based their decisions on the meaning of the given options and the context sentences. Interviewees shared their reflections on how they used meaning-based strategies to complete the test:

> *I looked at the full sentences first. In some sentences, I didn't have to do that because they were easier and quite obvious. For more difficult items, I had to look at the full sentences and the options.* (Vera NZ, Excerpt 5.25)

> *In the third test* [AC Recognition Test], *I read the context sentences and the options and chose the one that best fit the meaning.* (Thuong VN, Excerpt 5.26)

> *The options were given and if I knew the meaning of those, I could choose more quickly.* (Thu VN, Excerpt 5.27)

The key point about completing the AC Recognition Test was the understanding of the meaning of the academic collocations, given the comprehensible sentence prompts (see Section 5.3). Test-takers

only needed to pick the correct answer from the provided options, meaning the task would be very easy if they knew the meaning of the collocations which were presented. As long as the participants used this strategy to complete the test, the results would truly reflect their recognition knowledge of academic collocations, which in turn supported the Evaluation inference.

5.4.2.2 Priming effect

All the interviewees confirmed that they could recognise the resemblance between the AC Recognition Test and the AC Recall Test, although it took longer for some to realise this connection between the two tests than others. Because the two collocation tests use the same context sentences to elicit the same academic collocations, participants taking the two tests in sequence could still remember some aspects of the first test (i.e., the AC Recall Test) to do the latter test (i.e., the AC Recognition Test). Even with the VST (Nation & Beglar, 2007) including 140 test items as a filler task in between the two collocation tests, the priming effect might not be able to be completely ruled out because of the similarities between the tests. This carryover effect might or might not affect the validity of the AC Recognition Test results, depending on what aspects from memory the participants used to answer the test. The following five threads that emerged from the data were reported:

- General impression about the similarities between the tests
- Carryover memory of context sentences
- Carryover memory of correct responses
- Carryover memory of incorrect responses
- Carryover memory of initial letter hints

Firstly, not all the participants could articulate what they remembered specifically from one test to another. Eight interviewees seemed to only have a general impression about the similarities between the two collocation tests, as illustrated in Excerpts 5.28 and 5.29. This priming effect did not give any hints to the correct answers and therefore did not threaten the validity of the test results.

> *I just knew that the third test* [AC Recognition Test] *looked similar to the first test* [AC Recall Test] *and I found some words familiar, but I didn't remember in detail.* (Huyen VN, Excerpt 5.28)

> *I would say I could remember but not in detail. Some sentences when I read, I could recognise that I had seen them before. I could remember very few.* (Jin NZ, Excerpt 5.29)

Secondly, seven respondents revealed that they could remember the context sentences. The memory about the sentence prompts seemed to be just at the level of recognising that they had appeared before, as Jin reported in Excerpt 5.29. Khanh VN (Excerpt 5.30) is the only one who gave an estimate of his carryover memory. Even when the participants could remember precisely the sentences, the validity of the test results would not be threatened because the contexts contained no answers.

> *I could remember 50-60% of the context sentences repeated from the first test* [AC Recall Test] *to the third test* [AC Recognition Test]*, but I couldn't remember the phrases that I filled in.* (Khanh VN, Excerpt 5.30)

Thirdly, almost half of the interviewees (21/44) reported that they remembered some of their successful responses in the AC Recall Test. Participants knew whether their answers were correct or not for the first test when they saw the options in the AC Recognition Test. If a collocation that participants had filled in the AC Recall Test was among the options in the AC Recognition Test, they would be confident that their answer had been correct (see Excerpts 5.31 and 5.32).

> *I couldn't remember much. I only remembered some items in the first test* [AC Recall Test] *if I had put them right and I saw them again in the third test* [AC Recognition Test]. *Otherwise, even after I completed the third test* [AC Recognition Test]*, I still couldn't remember what the correct answers for the first test* [AC Recall Test] *were. It was hard to remember.* (Ngoc VN, Excerpt 5.31)

> *I did the third test* [AC Recognition] *by trying to recall the first test* [AC Recall Test]*, so it was affected to some extent. I could finish the last test very quickly. For example, in the first test* [AC Recall Test] *I filled in "urban area", then in the third test* [AC Recognition Test] *when I saw "urban area", I would choose it immediately without reading the whole sentence carefully. Because I had spent too much time thinking in the first test* [AC Recall Test]*, I was confident with my answer in the third test* [AC Recognition Test]. (Hong NZ, Excerpt 5.32)

It is not surprising that participants remembered correct answers in the AC Recall Test because they clearly knew the academic collocations they had recalled in the test. When test-takers were able to write down a correct collocation for the AC Recall Test, the chance that they would choose a correct option for the equivalent item in the AC Recognition Test was also high. However, if it is not assumed that recall knowledge implies recognition knowledge, the memory of correct answers in the AC Recall Test would pose a threat to the validity of the AC Recognition Test results.

Fourthly, 11 respondents reported having a memory of incorrect answers in the AC Recall Test. If participants did not see their answer in the listed options, they would believe that their answer in the previous test had been wrong (see Excerpt 5.33). Although this priming effect did not directly reveal the correct answers in the AC Recognition Test, it might direct participants to choose the option that shared the same initial letters with their incorrect response in the AC Recall Test (see Excerpt 5.35). However, not all the participants could recognise this connection, especially when the memory about incorrect responses seemed to be vaguer than the correct ones (see Excerpt 5.31).

> *I thought I had seen this question before. Then I tried to recall what I had put in for the first test* [AC Recall Test]. *When I saw the multiple-choice, I realised that what I had put in the first one* [AC Recall Test] *was wrong.* (Yun NZ, Excerpt 5.33)

The last point looks more closely at the carryover memory of the initial letter hints in the AC Recall Test. Nine respondents reflected on this aspect with comments and specific examples, such as:

> *When I did the first test* [AC Recall Test]*, I had to base my answers on the provided initial letter hints. When I came to the third test* [AC Recognition Test] *and couldn't select the answer, I would try to remember the initial letters in the first test* [AC Recall Test] *to choose the correct answer. But I could only remember the initial letters of some phrases; only a few.* (Hai VN, Excerpt 5.34)

> *I filled in "well done" in the previous test* [AC Recall Test]*, so it should be "well documented" now. I found "well done" a bit weird, so I could remember this one* [Item 10 on the AC Recognition Test]. (Nhi VN, Excerpt 5.35)

> *The initial letters "in..." and the word "value"* [she could only fill in the second word of Item 25 "intrinsic value" in the AC Recall Test] *reminded me of the word "invaluable", so I remembered that the adjective should start with "in...". That's why I chose B.* (Dieu VN, Excerpt 5.36)

When participants recognised the connection between the two collocation tests, they tried to recall the answers from the first test. However, as can be seen from Excerpt 5.34, remembering the initial letter hints was challenging. This might be because there were many test items, and the VST (Nation & Beglar, 2007) appeared between the two collocation tests. Respondents seemed to need some kind of association in order to remember the initial letter hints. For example, in Excerpt 5.35, Nhi VN remembered an incorrect answer to an item in the AC Recall Test, which prompted her to choose the option which shared the same initial letters for the same item in the AC Recognition Test. Dieu VN in Excerpt 5.36 gave another example when she associated the provided hints with a new word and therefore, she could remember the initial letters. Irrespective of how the respondents could remember the initial letter hints in the AC Recall Test, this carryover memory revealed the correct answers and thus affected the validity of the AC Recognition Test results.

The analysis so far has shown that there is evidence of the priming effect from the AC Recall Test to the AC Recognition Test. Nevertheless, participants did not report relying on their memory of the AC Recall Test as the main strategy for taking the AC Recognition Test. When being asked whether she did the AC Recognition Test on her own or remembered the answers from the AC Recall Test, Tran NZ said that:

> *It was a combination of both, but I would do it on my own first. Basically, the phrases were already in my head. I just needed to look at them again and I would remember and choose the correct ones.* (Excerpt 5.37)

In sum, depending on the aspects of the AC Recall Test that participants remembered, the validity of the AC Recognition Test results might or might not be affected. The general impression of having met some phrases or sentences before did not appear to cause any harm. The memory of the sentence prompts reduced the reading burden but did not reveal the answers, so the validity of the AC Recognition Test results should be unaffected. Recall of correct answers in the previous test indicated

that test-takers had knowledge of the collocations in the test and therefore their correct answers in the AC Recognition Test were expected, but still could cause a threat to the validity. Remembrance of wrong answers in the AC Recall Test revealed initial letter hints for the correct answer in the AC Recognition Test and thus might negatively affect the validity of the test results. That said, very little evidence was found concerning the ability of the participants to remember the initial letters. The section that follows looks into the last strategy employed by the test-takers for the AC Recognition Test.

5.4.2.3 Elimination and/or guessing

The strategies of elimination and guessing are widely reported as being integral parts of taking the multiple-choice test (Nevo, 1989; Salehi, 2011). Interviewees in this study revealed that they based their elimination on the meaning of the sentence prompt and the provided options. Participants seemed to think carefully for their answers even when they were uncertain. The following excerpts illustrate this point:

> *I read the context sentences and eliminated the most obvious wrong answers, then I chose the one with the most suitable meaning.* (Hong NZ, Excerpt 5.38)

> *If I didn't know the answer, I would use elimination. I would choose the option whose meaning was the most suitable. If the vocabulary size is large enough, the probability of choosing the correct answers will be higher.* (Trang NZ, Excerpt 5.39)

Random guessing was also employed by participants, as illustrated in Excerpt 5.40. This strategy differs from elimination in that it is not based on any type of knowledge. As a result, guessing has a lower chance of yielding the correct answer than using the elimination strategy.

> *I didn't want to leave an answer blank. So even though I didn't know, I still chose something.* (Anna NZ, Excerpt 5.40)

A correct answer based on elimination and/or guessing posed a threat to the validity of the AC Recognition Test results. This is because the answer did not truly reflect learners' knowledge of the

academic collocations tested. Despite not being a lucky guess, the correct answer using elimination indicated that participants had knowledge of the surrounding words but not the collocations tested.

Taken together, the strategies for the AC Recognition Test revealed by the respondents pointed out two issues affecting the validity of the test results. The first problem originated from the administration of the two collocation tests in a sequence, which caused the priming effect from the AC Recall Test to the AC Recognition Test. The second concern was related to the nature of the multiple-choice test in which test-takers provided an answer using elimination and/or guessing. That said, the main strategy by test-takers was still meaning-based, which indicated that the answers on the AC Recognition Test properly reflect learners' target knowledge.

This section reveals some aspects of the testing condition through test-takers' reflections on activities such as using dictionaries and the priming effect as a result of test administration. The following section will go into greater detail about how the interviews shed light on the participants' test-taking processes.

5.5 Test-takers' reflections on the test-taking processes

Information about the test-taking process can contribute to the evaluation of the testing condition and indicate whether there were any interruptions during the testing process. The completion time recorded by Qualtrics (i.e., the online testing platform) during test taking varied from almost an hour to 151 hours. I included a question in the interview to discover why there was such a significant difference in the test-taking time. The respondents were asked whether they finished the tests in one attempt, or they stopped and returned later (Question 8, Section B, Appendix B). It should be noted that Qualtrics allowed test-takers to save their progress. Among 44 interviewees, 20 participants explained that the length of the completion time was extended because they needed to change locations, or they had to stop to do other tasks. For example, Khanh VN said that she had to return to the tests three times to complete all three tests. Clearly, the interruptions in the test-taking time did not create an ideal testing environment and therefore threaten the Evaluation inference.

To discover whether the test battery was too long and might cause a fatigue effect, the average test-taking time was roughly estimated from the other 24 participants who reported that they finished the three tests in one attempt. The minimum time recorded by Qualtrics for these participants was 43

minutes and the maximum time was approximately two and a half hours. This included the time spent reading the consent form and completing the background questionnaire. On average, the participants completed the three tests in less than an hour and 45 minutes. English learners may already be familiar with language proficiency tests such as IELTS, TOEFL or TOEIC with a test-taking time of up to three to four hours. The roughly estimated completion time of the three tests in the current study of an hour and 45 minutes thus should be within the scope of practicality, meaning that the test length allowed the learners to demonstrate their vocabulary knowledge.

In sum, information about the test-taking time revealed by the respondents supported the Evaluation inference in terms of the appropriate test length, but there was also evidence against the test-taking condition with interruptions. Findings from this section confirm that the testing condition was not optimal for test-takers to demonstrate their knowledge of academic collocations (RQ2). The next section brings together the quantitative and qualitative evidence presented in this chapter to provide an overall assessment of the Evaluation inference in the validation framework of the ACTs.

5.6 Overall assessment of the Evaluation inference

This section discusses the extent to which the research findings support the Evaluation inference. Table 5.9 provides an overview of the warrants, evidence and judgement of the Evaluation inference. As shown in the table, the first warrant related to the characteristics of the test items was fully supported. The backing for this warrant was presented in Sections 5.2 to 5.4. The statistical analysis of the test items showed that overall, both the AC Recognition Test and the AC Recall Test conformed to the Rasch model. The 11 misfitting items in the AC Recognition Test and three misfits in the AC Recall Test could have been due to random variations in the statistical model. The findings from the analysis of unidimensionality and local independence indicated that the test items worked well together to measure the single construct of academic collocation knowledge. As for qualitative evidence, reflections from test-takers revealed that characteristics of the test tasks supported the elicitation of the target knowledge, as evidenced by the fact that the main strategies that participants employed to answer the tests were using prompts and meaning-based strategy. In brief, the first warrant concerning the test characteristics was fully supported.

Table 5.9

Summary of Warrants, Evidence and Degree of Support for the Evaluation Inference

Warrant	Evidence	Degree of support
The characteristics of the ACTs are appropriate to measure the intended construct of academic collocations.	• Statistical analysis overall showed appropriate fit of the ACT items to the Rasch model. • Reflections from test-takers revealed that the formats of the ACTs were appropriate to elicit knowledge of academic collocations.	Fully supported
The testing condition allows test-takers to demonstrate their knowledge of academic collocations.	Test-takers' reflections on the test-taking process revealed that: • Test length was appropriate. • Infrequent factors posed a threat to the validity of the test results: interruptions in the test-taking time, uses of dictionaries and priming effect resulted from the test administration.	Partially supported
The scoring procedure is appropriately developed and applied.	• The scoring method was carefully developed, piloted and explicitly documented in Chapter 4. • The scoring procedure was consistently applied as described in Chapter 4. • Post-test interview supported scoring blank responses as wrong answers.	Fully supported

The second warrant about the testing condition was partially supported. On the one hand, there was evidence that the test length was appropriate. On the other hand, data from the post-test interviews revealed factors that threatened the validity of the test results. First, the test-taking time of some participants was disrupted by other activities. External factors might hinder the ability of those test-takers to concentrate on the tests. Second, consulting dictionaries helped a few test-takers to answer some items in the AC Recall Test. This meant that the test did not elicit their own knowledge. Third, the priming effect caused by administering the two collocation tests in one sitting assisted participants in selecting the correct answer for some items in the AC Recognition Test. Responses based on memory from the earlier test did not precisely reflect learners' recognition knowledge of academic collocations. Although those factors counteract the Evaluation inference, they did not occur

frequently and thus may have only a minor impact on the test results. It is worth noting that all of those factors stemmed from the online test administration because of COVID-19, which caused a major limitation in this study (see Chapter 8, Section 8.3). Overall, it can be concluded that the second warrant of the Evaluation inference received only partial support.

The third warrant of the Evaluation inference concerning the scoring procedure was fully supported (see Table 5.9). The backing for this warrant was presented in Chapter 4. For the AC Recall Test, the scoring procedure was carefully developed, piloted and consistently applied. The lenient scoring method which accepted spelling and grammatical errors resulted in scores that reflected knowledge of academic collocations, and not the knowledge of written form and grammar. In addition, a response which was different from the target collocation was also accepted if its meaning fitted the context and it was also an academic collocation. The post-test interview provided further evidence to back up the scoring of blank answers. Participants revealed that they left items blank mostly because they had no knowledge of the tested items. Thus, it was completely reasonable that leaving answers blank was treated as wrong answers and resulted in zero points. As for the AC Recognition Test, the scoring was apparently objective and consistent, as automatically conducted by Qualtrics software. In sum, the backing provides complete support for the third warrant. Taken as a whole, all three warrants were supported to a certain extent, and thus the Evaluation inference was eventually upheld.

5.7 Chapter summary

The findings presented in this chapter provided both quantitative and qualitative evidence for the assessment of the Evaluation inference in the validation framework of the ACTs. This inference received partial support overall. The two warrants about the test item characteristics and the scoring were fully justified. The other warrant about the testing condition was partially supported. The next chapter will provide further validation evidence which focuses on the Generalisation and Extrapolation inferences. After that, the findings of these two chapters will be discussed together in Chapter 7.

Chapter 6 Validation results: Generalisation and Extrapolation inferences

This chapter continues to report on the validation results of the Academic Collocation Tests (ACTs). These findings serve as evidence to assess the Generalisation and Extrapolation inferences in the argument-based validation framework of the ACTs and answer the following research questions (RQ):

RQ3. Are scores on the ACTs reliable?
RQ4. Are scores on the ACTs related to scores on other tests measuring similar constructs?
RQ5. Is the item difficulty on the ACTs related to the frequency of academic collocations?
RQ6. Are scores on the ACTs related to English learning experience?

Section 6.1 reports the backing for the Generalisation inference, including internal consistency of the ACTs and test-retest reliabilities (RQ3). Section 6.2 then provides the evidence for the assessment of the Extrapolation inference, including correlations between the ACTs with other measures sharing similar constructs (RQ4), correlations between item difficulty of the ACTs and the frequency of academic collocations (RQ5), and correlations between the performance on the ACTs and English learning experience (RQ6). Finally, Section 6.3 summarises the findings of this chapter.

6.1 Validation results for the Generalisation inference

The Generalisation inference states that test-takers' scores on the ACTs are consistent across tasks and occasions. This is based on the warrant that the ACTs exhibited high internal consistency and high test-retest reliability. Different reliability indices are presented as evidence. First, Cronbach's alpha – the most commonly used measure of internal consistency and reliability indices from the Rasch model (Rasch, 1993) are reported. Then test-retest reliability indices are presented to assess the consistency of the test results across different occasions.

6.1.1 Cronbach's alpha and Rasch reliability of the ACTs

Both the AC Recognition Test and the AC Recall Test achieved high reliability irrespective of using Cronbach's alpha or Rasch reliability indices. The reliability information for these two tests is

presented in Table 6.1. The person reliability in the Rasch measurement analysis (Table 6.1, first row) can be interpreted similarly to Cronbach's alpha (Table 6.1, last row), meaning that a value closer to 1.00 indicates a higher internal consistency of a measure. Both the AC Recognition Test and the AC Recall Test achieved Cronbach's alpha value of .96, indicating very high reliability.

Table 6.1

Reliability Indices of the ACTs

	AC Recognition Test	AC Recall Test
Person reliability	.86	.94
Person separation	2.47	3.98
Item reliability	.96	.99
Item separation	4.89	9.18
Cronbach's alpha	.96	.96

However, Cronbach's alpha may overstate test reliability because it reports the reliability of raw scores that are sample-dependent, while Rasch reliability is less misleading for inference beyond the sample (Linacre, 1997). The AC Recognition Test with the person reliability of .86 can discriminate the person sample into two or three levels, while the AC Recall Test with the reliability of .94 can divide the sample into three or four levels (see Linacre, 2019). The person separation indices of 2.47 for the AC Recognition Test and 3.98 for the AC Recall Test show that both tests are sensitive enough to distinguish between high and low performers. The item reliability indices (above .90) and the item separation indices (above 3.00) of the two tests imply that the person sample was large enough to confirm the item difficulty hierarchy of the instrument (Linacre, 2019).

6.1.2 Test-retest reliability of the ACTs

Both the AC Recognition Test and the AC Recall Test were found to have high test-retest reliability. Test-retest reliability was calculated to investigate how the tests could reliably replicate the results when being administered with the same person sample on different occasions. It is important to remember that 44 test-takers who completed the online ACTs sat the same tests again verbally in post-test interviews within two weeks after the first administration. In the present study, the test-retest reliability was indicated by Spearman's correlation coefficient (r_s) which measures the strength of the relationship between the two data sets. If the tests could consistently produce the same results, then the relationship between the two score sets (pre and post) should be high. Table 6.2 reports the

summary of participants' results of the online tests (i.e., 'Test' columns) and results of the verbal tests (i.e., 'Retest' columns).

Table 6.2

Summary of ACT Test-Retest Results (N = 44)

	AC Recognition		AC Recall	
	Test	Retest	Test	Retest
Mean	51.59	53.39	33	33.89
SD	8.02	7.01	11.47	12.67
r_s ($p < 0.01$)	**.86**		**.87**	

As shown in Table 6.2, the mean scores of the ACTs are higher in the 'Retest' columns compared to the 'Test' column, indicating that participants tended to score slightly higher when they retook the ACTs the second time. The findings showed significantly strong relationships between test and retest results in both the AC Recognition Test ($r_s = .86$, $p < 0.01$) and the AC Recall Test ($r_s = .87$, $p < 0.01$), indicating that scores on the ACTs produced consistent results across different occasions. Overall, the findings provide the answer to RQ3, thereby confirming that the ACTs were highly reliable.

6.1.3 Overall assessment of the Generalisation inference

Table 6.3 summarises the warrants, evidence and overall judgement of the Generalisation inference. As can be seen from the table, the first warrant about test reliability was fully supported with high Rasch reliability and high Cronbach's alpha. The second warrant, concerning the test-retest reliability, also received full support with strong correlations of test results between two testing occasions. Taken as a whole, it can be concluded that the Generalisation inference stands.

Table 6.3

Summary of Warrants, Evidence and Degree of Support for the Generalisation Inference

Warrants	Evidence		Degree of support
	AC Recognition Test	AC Recall Test	
The ACTs exhibit high reliability.	• High person reliability (.86) • High person separation (2.47) • High item reliability (.96) • High item separation (4.89) • High Cronbach's alpha (.96)	• High person reliability (.94) • High person separation (3.98) • High item reliability (.99) • High item separation (9.18) • High Cronbach's alpha (.96)	Fully supported
Scores on the ACTs produce consistent results across occasions.	• High test-retest reliability (.86)	• High test-retest reliability (.87)	Fully supported

6.2 Validation results for the Extrapolation inference

The Extrapolation inference states that performance on the ACTs is indicative of the target construct of academic collocations. This inference relies on three warrants: 1) scores on the ACTs are related to scores on other tests that measure similar constructs; 2) item difficulty on the ACTs is related to the frequency of academic collocations, and 3) scores on the ACTs are related to English learning experience. First, correlations between tests with related constructs are presented as backing. Second, correlations between test item difficulty and the frequency of academic collocations are reported. Finally, correlations between scores on the ACTs and English learning experience add further evidence for the Extrapolation inference.

6.2.1 Correlations between tests of similar constructs

Scores on tests of similar constructs are expected to have positive relationships, including the following:

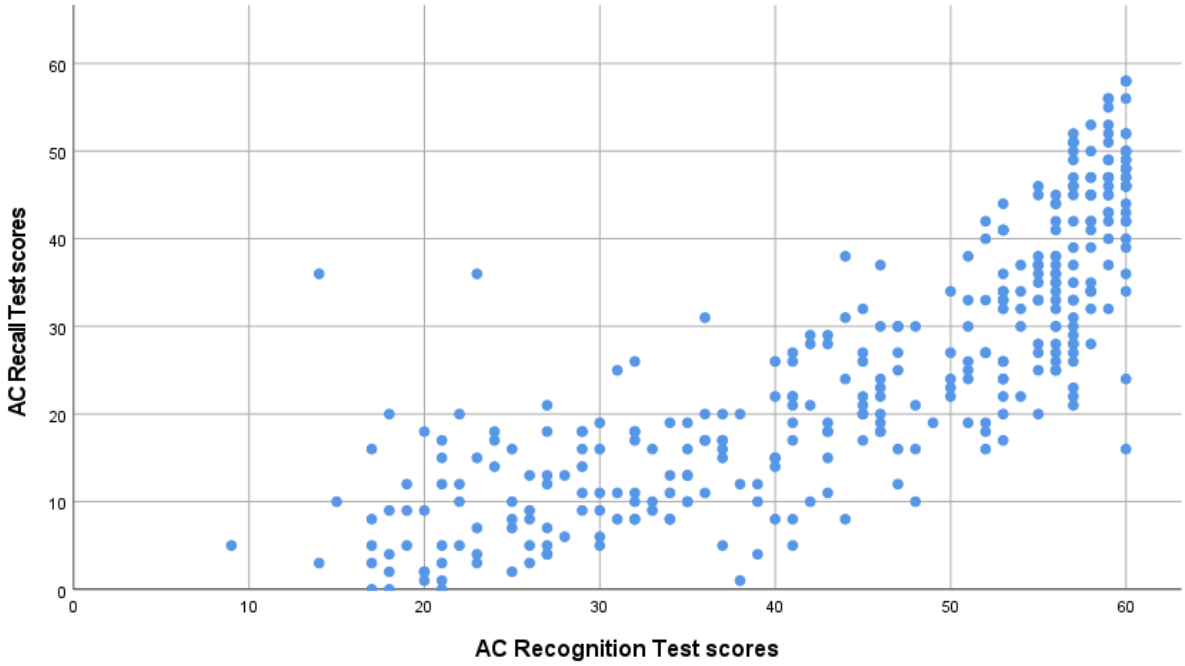
- The AC Recognition Test and the AC Recall Test, both of which measure knowledge of academic collocations
- The ACTs and the Vocabulary Size Test (VST) (Nation & Beglar, 2007), both of which measure knowledge of English vocabulary
- The ACTs and English language proficiency tests (e.g., IELTS or TOEFL), both of which measure knowledge of English language

6.2.1.1 Correlations between scores on the AC Recognition Test and the AC Recall Test

The AC Recognition Test and the AC Recall Test were found to have a strong positive correlation. As can be seen from the scatterplot in Figure 6.1, the AC Recall Test scores tended to increase when the AC Recognition Test scores grew. The strength of correlation between the two score sets of the ACTs was confirmed by the correlation coefficient r_s = .86 (p <0.01). As both the AC Recognition Test and the AC Recall Test measure knowledge of academic collocations, the high correlation between them supports the Extrapolation inference.

Figure 6.1

Correlation Between Scores on the AC Recognition Test and the AC Recall Test



6.2.1.2 Correlations between scores on the ACTs and the VST (Nation & Beglar, 2007)

There were significant relationships between scores on the ACTs and the VST. As illustrated by Figures 6.2 and 6.3, participants with greater vocabulary size tended to score higher on the ACTs, indicating the positive relationship between general vocabulary knowledge and knowledge of academic collocations. The correlation coefficients confirm the expectation showing that scores on the AC Recognition Test (r_s = .53, p < 0.01) and the AC Recall Test (r_s = .52, p < 0.01) moderately correlated with scores on the VST. Since both the ACTs and the VST measure knowledge of English vocabulary, the correlations found between them give support for the Extrapolation inference.

Figure 6.2

Correlation Between Scores on the VST and the AC Recognition Test

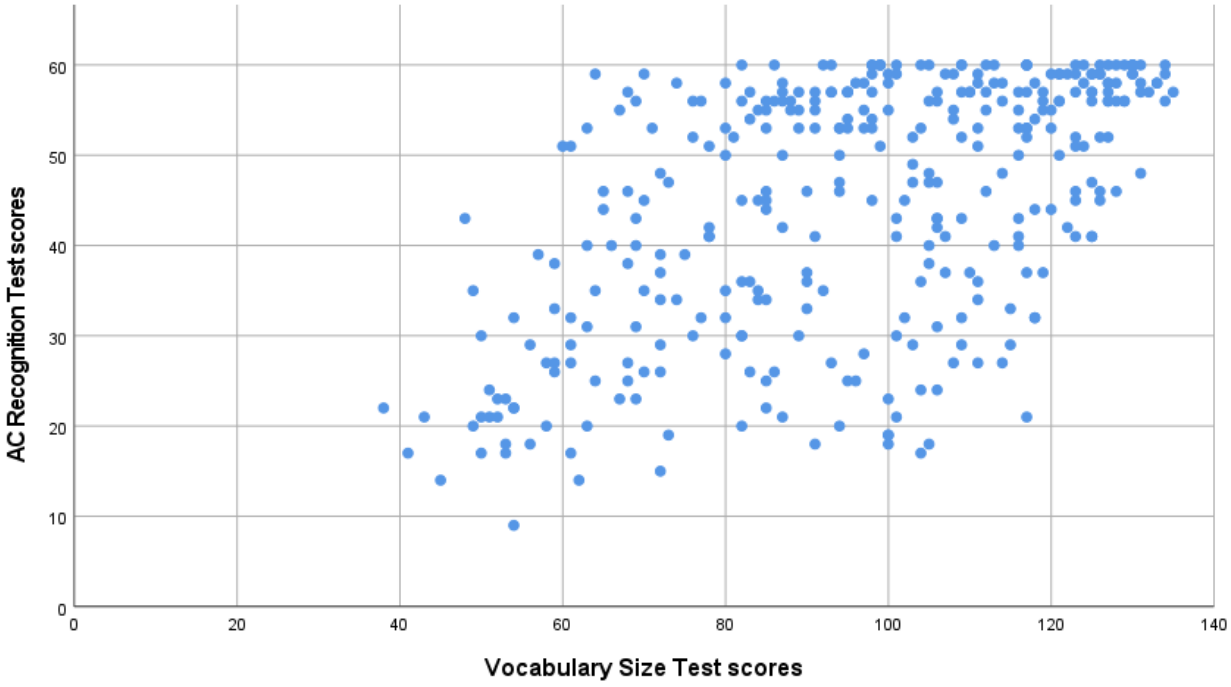
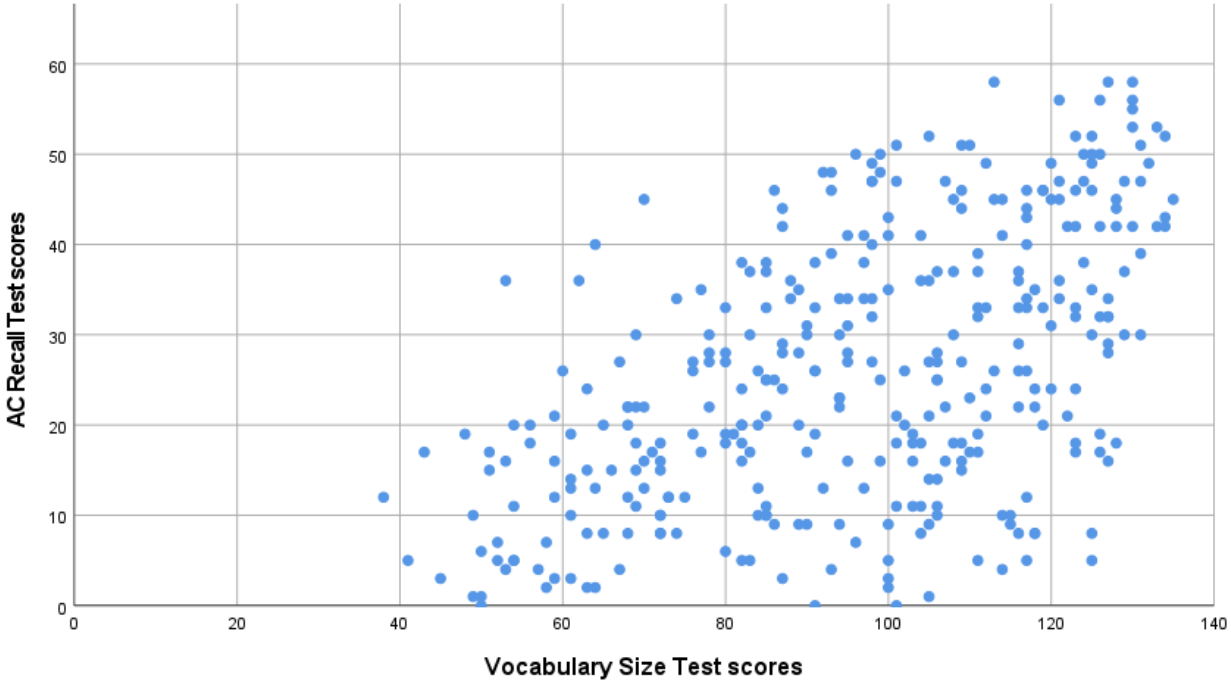


Figure 6.3

Correlation Between Scores on the VST and the AC Recall Test



6.2.1.3 Correlations between scores on the ACTs and English language proficiency tests

Those who attend university in New Zealand have to provide proof of English language proficiency as one of the entrance requirements. The evidence includes a degree in a program with English as the language of instruction in an English-speaking country or a report of an English language proficiency test (e.g., IELTS or TOEFL). Among 110 participants in New Zealand, 82 provided their IELTS scores and eight reported their TOEFL scores in the background questionnaire. IELTS and TOEFL assess four English language skills (i.e., listening, reading, writing and speaking), but use different scoring scales. The maximum band score for IELTS is 9.0 and for TOEFL it is 120. Before calculating the correlations, the TOEFL scores were converted to the equivalent IELTS band scores following the official conversion page of the Educational Testing Service (available at https://www.ets.org/toefl/score-users/scores-admissions/compare/). Table 6.4 summarises the IELTS results of 90 participants in New Zealand, including 82 official IELTS scores and eight converted scores from TOEFL.

Table 6.4

Summary of New Zealand Participants' IELTS Scores (N = 90)

	Min	Max	Mean	SD
IELTS scores	5.50	9.00	7.27	0.65

The correlation results indicated positive relationships between scores on the ACTs and IELTS. To be more specific, scores on the AC Recognition Test showed a modest correlation with the IELTS scores (r_s = .28, p = 0.009), while scores on the AC Recall Test moderately correlated with the IELTS scores (r_s = .53, p < 0.01). As illustrated by Figures 6.4 and 6.5, the scatter plot is more dispersed in Figure 6.4, indicating that the relationship is weaker in this pair of data sets.

Figure 6.4

Correlation Between Scores on the IELTS and the AC Recognition Test

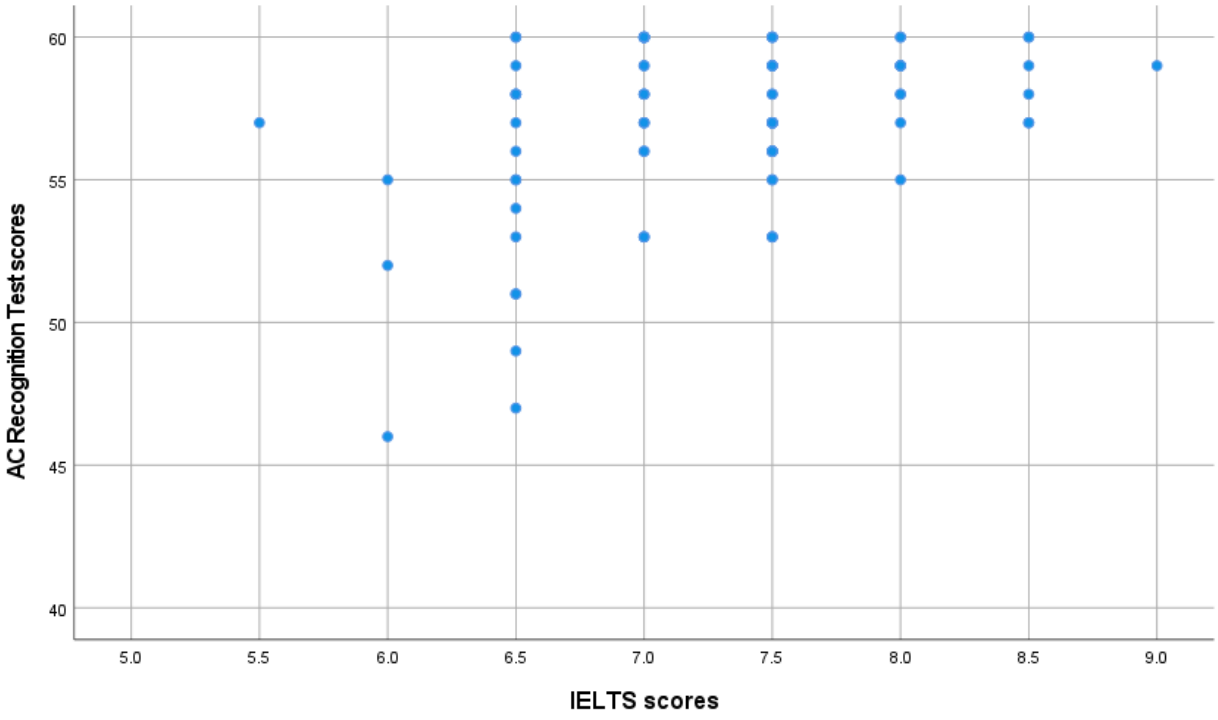
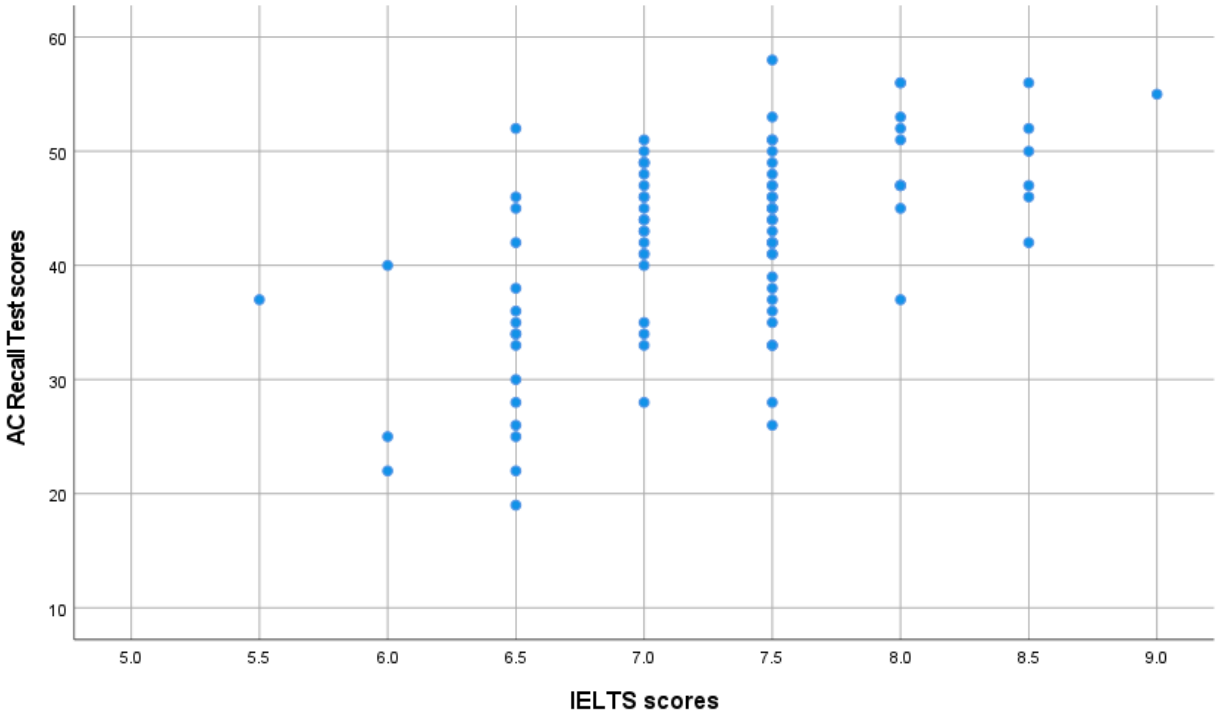


Figure 6.5

Correlation Between Scores on the IELTS and the AC Recall Test



The correlations between scores on the ACTs and the IELTS were not strong, and this finding was, to a certain extent, to be expected. Vocabulary is not directly measured through these tests but is incorporated into different skills-based test sections. For example, receptive knowledge of vocabulary is assessed through reading and listening, while productive knowledge of vocabulary is assessed via speaking and writing. The ACTs, on the other hand, focus specifically on a special type of vocabulary, that is, academic collocations. Moreover, the IELTS scores recorded in the background questionnaire could have been outdated as the participants had taken the test before their official entrance to the university in New Zealand. For example, one participant was a final-year PhD candidate who had taken the IELTS test four years before the current study took place. This means the IELTS scores might not truly reflect learners' language proficiency level at the time they took the ACTs. That said, if participants' IELTS scores were employed, the positive correlations between scores on the ACTs and the IELTS scores would provide backing for the Extrapolation inference.

In brief, the results presented in this section provide the answer to RQ4, confirming that scores on the two ACTs were strongly correlated and they were both related to scores on other tests of similar constructs, including the VST and the IELTS.

6.2.2 Correlations between item difficulty of the ACTs and the frequency of academic collocations

Further correlation analysis helps to test an assumption that the test items containing more frequent academic collocations will be easier than items using less frequent collocations. In other words, the number of correct answers would be higher on test items that were based on more frequent collocations. Frequency information of academic collocations and the number of correct answers for each test item are in Appendix G. The range of item frequency is from 112 to 7,220 occurrences in 112 million words. Although the test items were selected from a corpus-based word list of the most frequent academic collocations, the frequency might seem to be low as collocations are not as frequent as single words. As shown in Figures 6.6 and 6.7 (or Appendix G), there are only five items with a frequency above 1,000.

Figure 6.6.

Correlation Between Corpus Frequency of Academic Collocations and Item Difficulty of the AC Recognition Test

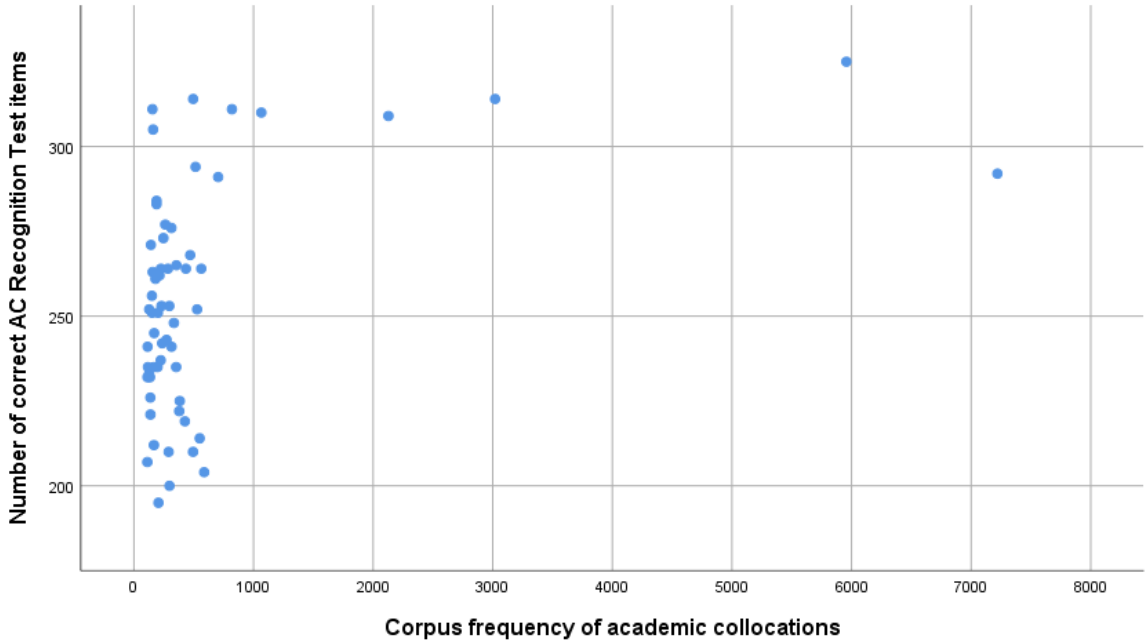
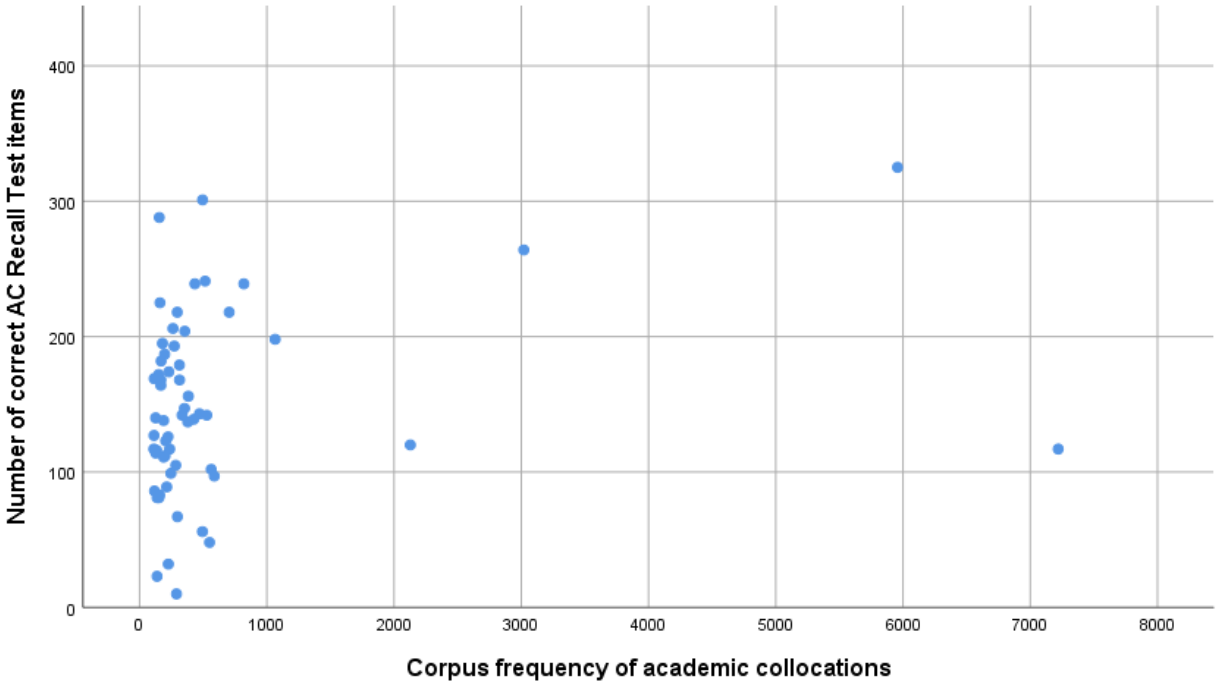


Figure 6.7

Correlation Between Corpus Frequency of Academic Collocations and Item Difficulty of the AC Recall Test



The results showed a positive correlation between the frequency and the item difficulty in the AC Recognition Test (r_s = .34, p < 0.01) and a statistically non-significant correlation in the AC Recall Test (r_s = .24, p = 0.06). The findings are illustrated in Figures 6.6 and 6.7. Although the majority of the items have a frequency lower than 1,000, the larger values seem to be influential and dictate the scale of the plot. To provide a clearer picture, Figures 6.8 and 6.9 illustrate the correlation without five items with a frequency greater than 1,000. The spots seem to be a random mix in these two scatterplots in that they show almost no relationship. The correlation results when removing the five most frequent items were not significant with r_s < .18, p > 0.1. Consequently, there seems to be no significant correlation between corpus frequency and the difficulty of test items in the ACTs, which answers the RQ5. This finding does not support the warrant that item difficulty of the ACTs is related to the frequency of academic collocations.

Figure 6.8

Correlation Between Corpus Frequency of Academic Collocations Excluding Five Items With a Frequency Above 1,000 and Item Difficulty of the AC Recognition Test

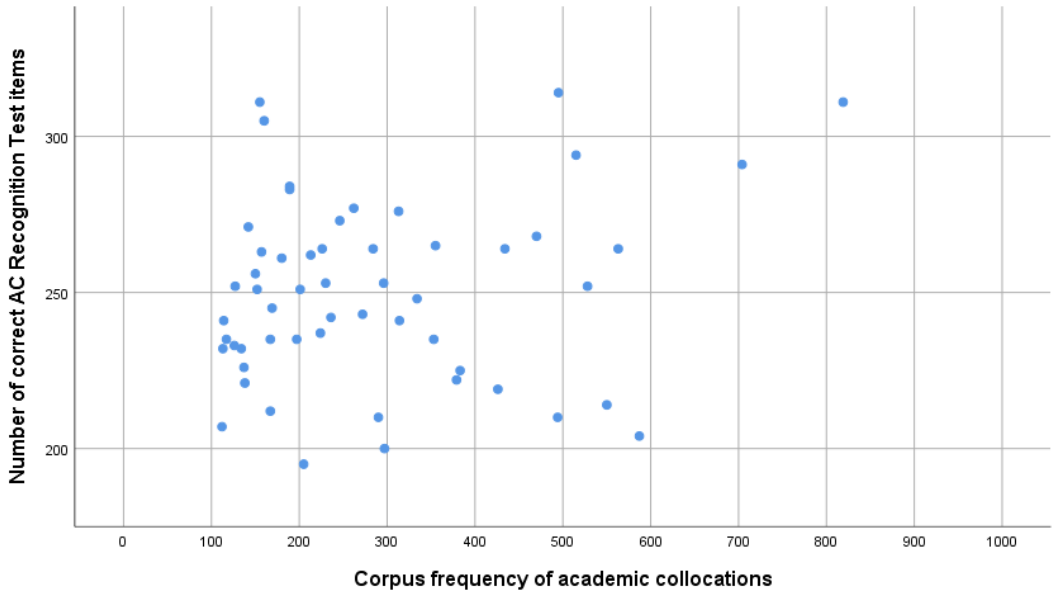
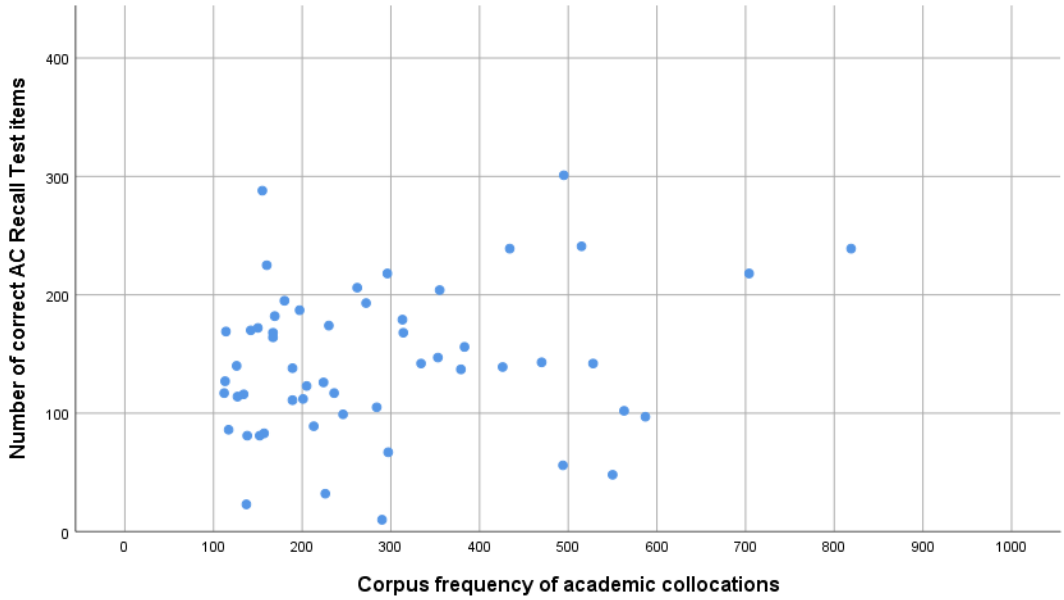


Figure 6.9

Correlation Between Corpus Frequency of Academic Collocations Excluding Five Items With a Frequency Above 1,000 and Item Difficulty of the AC Recall Test



6.2.3 Correlations between scores on the ACTs and English learning experience

Learners with more English learning experience are expected to have better knowledge of academic collocations. To examine this hypothesis, in the background questionnaire, respondents were asked to report the total time they studied English and the total time they studied in an English-speaking country. Correlation results between these factors and scores on the ACTs are reported below.

6.2.3.1 Correlations between scores on the ACTs and time of studying English

Time of studying English included the total time participants learned and used English for academic purposes in English as a Foreign Language (EFL) and English as a Second Language (ESL) contexts. Table 6.5 summarises the time of studying English reported by both EFL students in Vietnam and ESL participants in New Zealand. On average, the respondents had more than 12 years of learning English at the time of taking the ACTs.

Table 6.5

Participants' Time of Studying English in Years

	N	Min	Max	Mean	SD
EFL participants	233	5	19	10.29	2.46
ESL participants	110	2	30	17.12	6.97
Total	**343**	**2**	**30**	**12.47**	**5.45**

Participants' time of studying English has a moderate correlation with their scores on both the AC Recognition Test (r_s = .39, p < 0.01) and the AC Recall Test (r_s = .43, p < 0.01), as illustrated in Figures 6.10 and 6.11. These results provide further evidence to support the Extrapolation inference.

Figure 6.10

Correlation Between Years of Studying English and Scores on the AC Recognition Test

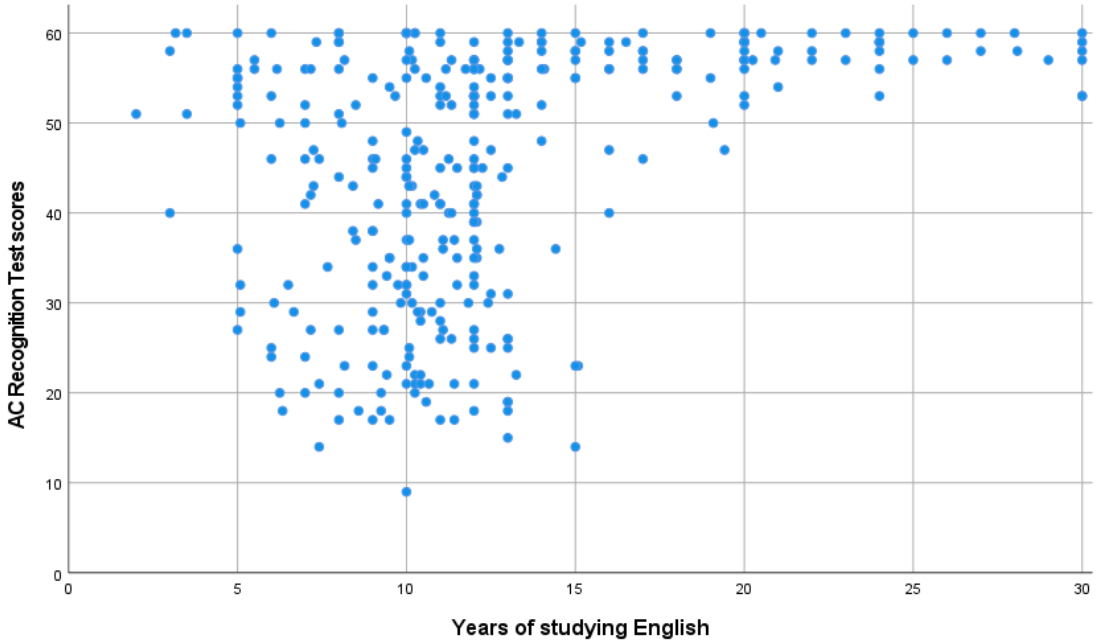
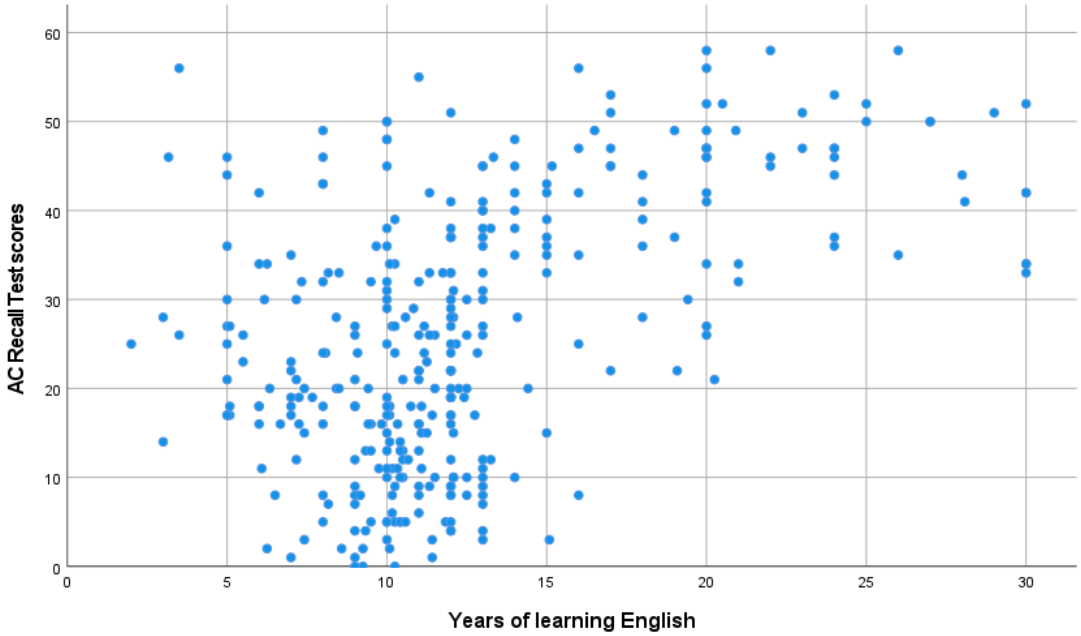


Figure 6.11

Correlation Between Years of Studying English and Scores on the AC Recall Test



6.2.3.2 Correlations between ACT scores and time of studying in an English-speaking country

Only participants in New Zealand reported the time they studied in an English-speaking country (see Table 6.6). The time could include the period they attended another program outside New Zealand. For example, a third-year PhD student in New Zealand who had spent two years doing her MA in Australia reported the total time of five years learning in English-speaking countries. The average time recorded for all the ESL participants was around four years.

Table 6.6

ESL Participants' Time of Studying in an English-Speaking Country in Years

N	Min	Max	Mean	Median	SD
110	1.0	27	4.29	3.0	4.62

The results showed statistically non-significant correlations between the time of studying in an ESL context with scores on both the AC Recognition Test (r_s = .02, p = .88) and the AC Recall Test (r_s = .15, p = .11). The lack of correlation between the two variables indicated that the ESL learning context did not affect learners' knowledge of academic collocations to a significant degree. This finding does not support the Extrapolation inference.

One question that emerged as a result of the lack of ESL context correlation was whether knowledge of academic collocation was different between ESL and EFL students. Table 6.7 summarises the test results of the EFL group (i.e., participants in Vietnam) and the ESL group (i.e., participants in New Zealand).

Table 6.7

Summary Statistics on the Test Scores of the EFL and ESL Groups

		AC Recognition Test		AC Recall Test		VST	
	N	Mean	SD	Mean	SD	Mean	SD
EFL group	233	38.82	13.07	18.95	10.99	90.33	23.96
ESL group	110	56.75	3.67	41.04	9.90	107.73	17.74

Results of independent t-tests showed that the ESL group (M = 56.75, SD = 3.67) demonstrated significantly better scores in the AC Recognition Test, $t(341) = -14.13$, $p < .001$, $d = 1.63$ compared to the EFL group (M = 38.82, SD = 13.07). Similarly, the ESL group (M = 41.04, SD = 9.90) scored

significantly higher than the EFL group (M = 18.95, SD = 10.99) on the AC Recall Test, $t(341) = -17.92$, $p < .001$, $d = 2.07$. Taken together, there was a significant difference in scores on the ACTs between the EFL group and the ESL group. The better knowledge of the ESL group could not be attributed to the learning context because participants' scores on the ACTs had no significant correlation with the time they spent studying in an English-speaking country.

Then the remaining question is what factors led to the difference in the ACT scores between the EFL and the ESL groups. One hypothesis is that the group of ESL participants scored higher on the ACTs because their language proficiency was higher overall than the EFL. The results on the VST support this hypothesis. The ESL group (M = 107.73, SD = 17.74) scored significantly higher on the VST compared to the EFL group (M = 90.33, SD = 23.96), $t(341) = -6.79$, $p < .001$, $d = 0.79$ (Table 6.7). Additionally, the average IELTS score of the ESL group was above 7.0 (see Table 6.4). This score is an indication of their advanced level of English proficiency. In general, an IELTS score between 4.5 and 7.0 is required for students in Vietnam to complete their degree (British Council, n.d.; Trang, 2010). This means that the current students usually have not yet reached the level of IELTS 7.0 because they have not finished their studies.

Altogether, the ESL context was not the factor that affected the acquisition of academic collocations. The fact that the ESL participants had higher language proficiency levels than the EFL counterparts might contribute to the difference in the ACT scores between the two groups. This finding does not provide backing for the Extrapolation inference. In sum, the answer to RQ6 is that scores on the ACTs were related to time spent learning English but not time spent studying in an English-speaking country.

6.2.4 Overall assessment of the Extrapolation inference

Table 6.8 summarises the warrants, evidence and overall judgement of the Extrapolation inference. The first warrant on the relationship between ACT scores and other tests of similar constructs was fully supported, as the findings showed that the ACTs were positively correlated with the VST and IELTS. The AC Recognition Test and the AC Recall Test also had a strong correlation. The second warrant is not supported since there was no evidence of a correlation between test item difficulty and corpus frequency. The third warrant was partially supported as the time of studying English was

moderately correlated with scores on the ACTs, although the time in an ESL context had no correlation with the test scores.

Table 6.8

Summary of Warrants, Evidence and Degree of Support for the Extrapolation Inference

Warrants	Evidence		Degree of support
	AC Recognition Test	AC Recall Test	
Scores on the ACTs are related to scores on other tests that measure a similar construct.	• High correlation with the AC Recall Test (r_s = .86) • Moderate correlation with the VST (r_s = .53) • Weak correlation with the IELTS (r_s = .28)	• High correlation with the AC Recognition Test (r_s = .86) • Moderate correlation with the VST (r_s = .52) • Moderate correlation with IELTS (r_s = .53)	Fully supported
Item difficulty on the ACTs is related to the frequency of academic collocations.	• Non-significant correlation with corpus frequency	• Non-significant correlation with corpus frequency	Not supported
Scores on the ACTs are related to English learning experience.	• Moderate correlation with time studying English (r_s = .39) • Non-significant correlation with time spent in an ESL context	• Moderate correlation with time studying English (r_s = .43) • Non-significant correlation with time spent in an ESL context	Partially supported

Taken altogether, two out of the three warrants were supported; therefore, the Extrapolation inference held.

6.3 Chapter summary

The results presented in this chapter provided evidence for the warrants in the Generalisation and Extrapolation inferences. The Generalisation inference was fully supported with high reliability and high test-retest reliability of the ACTs as evidence. The Extrapolation inference was partially supported. There was a strong correlation between scores on the AC Recognition Test and the AC Recall Test, and the scores on both tests positively correlated with scores on the VST (Nation & Beglar, 2007) and the IELTS. Corpus-based frequency did not seem to be a strong predictor of the

acquisition of academic collocations as the frequency of academic collocations did not have a relationship with the test item difficulty. Finally, scores on the ACTs moderately correlated with the time studying English, but not the time in an ESL context. In Chapter 7, these findings will be drawn together with the findings from Chapter 5 to provide a bigger picture of the development of academic collocation knowledge (Section 7.1). Chapter 7 will also discuss the inferences in the validity argument of the ACTs in greater depth, including the most important inferences and the impact of an inference not being fully supported (Section 7.3).

Chapter 7 Discussion

The overarching aim of the present study is to develop and provide validity assessment for the Academic Collocation Tests (ACTs). The current chapter focuses on three main issues that have arisen from this thesis so far. They are: (1) developing academic collocation knowledge, (2) using word lists in creating tests of academic collocations, and (3) applying development and validation frameworks for academic collocation tests. These three themes are missing pieces in the research literature. The first theme contributes to a new understanding in vocabulary studies. The last two themes are related to new applications of available resources to advance the field of vocabulary testing. Let us look at these themes in turn.

7.1 What conclusions can be drawn about the development of academic collocation knowledge based on the results of the ACTs?

The process of testing EFL/ ESL learners' recognition and recall knowledge of academic collocations in this thesis has paved the way to finding out more about the nature of the development of this knowledge. Although a complete picture of the acquisition of academic collocations cannot be captured within the scope of this thesis, the following five points provide an initial sketch of learners' knowledge of this lexis:

- EAP students tend to have substantial recognition knowledge but much less recall knowledge of academic collocations.
- Recognition and recall knowledge of academic collocations are strongly correlated.
- Knowledge of academic collocations is in turn positively correlated with learners' vocabulary size, their general language competence and years of studying English.
- Frequency might not be a strong factor in predicting the ease of learners' acquisition of knowledge of academic collocations.
- Time spent in an ESL context has a non-significant relationship with academic collocation knowledge.

These points are elaborated on further below.

7.1.1 Learners' recognition and recall knowledge of academic collocations

The present study is the first that directly compares learners' recognition and recall knowledge of the same target academic collocations. The findings of this study suggest that the EFL/ESL participants tended to have substantial recognition knowledge but much less recall knowledge of academic collocations. On average, test-takers scored almost 75% on the AC Recognition Test and less than 45% on the AC Recall Test. Studies on general collocation knowledge reveal that learners have limited ability to both recognise and recall collocations (e.g., Bahns & Eldaw, 1993; Gitsaki, 1999; Macis & Schmitt, 2017; Nguyen & Webb, 2017). The current study, however, indicates that recognising academic collocations is not a major issue but recalling them is. This finding is in line with Voss (2012) and Wongkhan and Thienthong (2020). Voss (2012) found the average scores on his academic collocation test were below 20%, meaning that his participants demonstrated very limited recall knowledge of collocations. Meanwhile, participants in Wongkhan and Thienthong's (2020) study showed good recognition knowledge of academic collocations as they tended to select the most appropriate word that frequently co-occurred with the node word.

The fact that test-takers achieved high scores on the AC Recognition Test is expected. As an academic collocation is usually transparent, learners can possibly infer its meaning if they know the meaning of its components (e.g., *significant difference*). It is likely that the participants would have known the individual words of the academic collocations. General academic vocabulary tends to occur within the first three to four thousand word families of the BNC/COCA lists (Nation, 2016). The Vocabulary Size Test (VST) (Nation & Beglar, 2007) results showed that the participants had an average vocabulary size of above 9,000 word families, which means these words are likely to be known or recognised in the AC Recognition Test. Furthermore, the possibility that providing distractors with one word in common (e.g., *academic achievements* and *academic performances*) cannot be ruled out as a cause for inflation of test-takers' scores. In other words, the high scores in the AC Recognition Test might partly be due to the highly controlled test format used in the present study. It should also be noted that the AC Recognition Test is perhaps not too easy but just easy for this group of participants who included English majors and those who had been studying in New Zealand for a while.

There are several possible reasons for participants' lower AC Recall test scores. First of all, data from post-test interviews show that test-takers tended to think about individual words and synonyms when doing the AC Recall Test, meaning that the academic collocations were not formed as holistic units in those participants' lexicon. For example, learners produced malformed collocations such as *draw a discrepancy* instead of *draw a distinction* or *internal value* instead of *intrinsic value*. These deviant academic collocations, which are unnatural word combinations to native and high proficiency English language speakers rather than completely nonsense combinations, could be the result of learners being unaware that synonyms often require different collocates (Liu & Zhong, 2016).

The second possible reason is that learners have not encountered the target academic collocations frequently enough to establish them in their memories (Hoey, 2005; Schmitt, 2013). According to Nation (2013), academic collocations may not be frequent enough to be acquired implicitly, nor are they parts of the technical vocabulary that may be taught explicitly. Textbooks might be among the main sources that provide exposure to academic collocations for learners. However, research on EAP textbooks has found deficiencies in exposure to academic collocations (e.g., Coxhead et al., 2020; Vu & Michel, 2021). This vocabulary seems not to receive adequate attention compared to other aspects such as grammatical functions in English classes (Begagić, 2015; Dokchandra, 2019).

Finally, the test format could also play a role in the low scores on the AC Recall Test. The initial-letter-hint restricted the possible answers for test items. When sharing opinions about this hint in interviews, respondents revealed that they could occasionally think of other collocations to fill in the gaps but failed to provide the one whose initial letters matched the prompts. Unfortunately, these participants could not provide any specific examples to illustrate their point.

In brief, the current study indicates that academic collocations are usually not very difficult to recognise, but they can be challenging to recall correctly. Given the participants' poor performances on the AC Recall Test, this thesis advocates for greater attention to academic collocations, especially in EAP textbooks and programs (see Chapter 8).

7.1.2 Strong correlation between recognition and recall knowledge of academic collocations

Another finding from this study is that recognition and recall knowledge of academic collocations are strongly correlated (r_s = .86), which indicates that they are two distinct but closely related constructs.

This finding corresponds to Fernández and Schmitt (2019) who reported a strong correlation (r_s = .81) between recognition and recall knowledge of general collocations. Fernández and Schmitt (2019) stress that the distinction between recognition and recall mastery is the most critical aspect in vocabulary knowledge. The present study highlights the importance of having separate tests to measure recognition and recall knowledge, as suggested by Webb (2005), in order to have a more accurate indication of learners' academic collocation knowledge. If only one test is used for different testing purposes (e.g., diagnosing learners' academic collocation knowledge for English reading and writing courses), the results could be misleading at worst and provide an incomplete picture at best.

7.1.3 Factors that positively correlated with knowledge of academic collocations

Vocabulary size, language proficiency and academic experience are interrelated. It is not surprising that if collocation knowledge is correlated with one element, it will also be associated with the other two. The results of this study suggest that learners with greater vocabulary knowledge, as indicated by scores on the VST (Nation & Beglar, 2007), are more likely to recognise correct academic collocations and produce acceptable ones, although this relationship is just moderate (r_s > .50). Previous studies on knowledge of general collocations (e.g., Fernández & Schmitt, 2019; Gyllstad, 2009; Nguyen & Webb, 2017) reported a positive but stronger correlation with vocabulary size (r > .70). As a vocabulary size test measures knowledge of general vocabulary, a higher correlation with the tests on general collocations than with the ACTs is expected. A correlation between the ACTs and a test of academic vocabulary would thus be predicted to be stronger.

This study also found a positive relationship between scores on the ACTs and IELTS as a language proficiency test. Interestingly, the correlation with the AC Recall Test was higher (r_s = .53) than with the AC Recognition Test (r_s = .28). The result suggested that the recall test might predict learners' general proficiency better than the recognition test. Additionally, this study discovered a moderate relationship between academic collocation knowledge and years of learning English, with a correlation coefficient r_s of about .40. This finding is in line with Frankenberg-Garcia (2018) and Wongkhan and Thienthong (2020) who found that participants with more academic experience had better knowledge of academic collocations than those with less experience. Similarly, Fernández and Schmitt (2015) reported a moderate correlation (r = .45) between years of study and collocation knowledge. Overall, the present study suggests that greater vocabulary size, higher language

proficiency and more English learning experience will help learners better acquire academic collocations.

7.1.4 Negligible effect of frequency on knowledge of academic collocations

Looking further into the development of academic collocations, this study suggests that, unlike single words, frequency does not appear to be a predictor of academic collocation knowledge. Participants in the current research did not tend to know high-frequency academic collocations better than the less frequent ones. For example, the collocation *driving force*, which was recognised by almost 60% of the test-takers but correctly recalled by only 28% of them, is more frequent than *solar power* (587 to 155 occurrences in the COCA Academic corpus), which was recognised by nearly 91% of the test-takers and correctly recalled by about 84%. This finding is consistent with Macis and Schmitt (2017) and Voss (2012) who reported non-significant frequency effect. However, Fernández and Schmitt (2015) and Chen (2019) found that frequency positively correlated with the number of correct answers for each collocation test item.

The difference in those research findings regarding the effect of frequency on collocation knowledge can be explained in several ways. First, corpus frequency counts may vary depending on how the collocations are defined and identified in each study. In the present study, the frequency information was calculated based on two-word collocations (see Chapter 4, Section 4.3). The frequency of collocations changes according to the span of a frequency search (i.e., how many words to the left or to the right of the node word). For instance, the frequency of the collocation *meet criteria* on the COCA Academic corpus is 203 occurrences with the span of searching one word to the right, and this frequency increases to 646 when the span is expanded to two words to the right (e.g., *meet the criteria*). This difference in setting the span and calculating the frequency of collocations may lead to the discrepancy in the correlation results.

Second, it is possible that the lack of correlation between frequency and knowledge of academic collocations in this study stems from the fact that the frequency of academic collocations is generally low. Although the academic collocations in the current research belong to the group of highly frequent collocations (Ackermann & Chen, 2013), there are only five out of 60 items (i.e., *significant difference, physical activity, wide range, address (an) issue* and *religious belief*) with a frequency

above 1,000 in the COCA Academic corpus of 111 million words. The other 55 items have a frequency ranging from 819 to 112 in the same corpus. The overall frequency of collocations is much lower than that of single words. For example, *significant difference* occurs 7,220 times in the COCA Academic corpus. It is one of the most frequent collocations in the ACT test items. The frequency of *significant* and *difference* as individual words in the same corpus is much higher: 60,470 occurrences for *significant* and 63,990 for *difference*.

Third, there is not much difference between the frequency of test items from different frequency bands in the current research. Although the academic collocations from the list of Ackermann and Chen (2013) were divided into ten frequency bands for even sampling, the frequency of the collocations in these bands was very close. For example, the collocation *specific type* with the raw frequency of 382 in the COCA Academic corpus belongs to Band 3, while the collocation *make (a) transition* with a frequency of 380 in the same corpus is in Band 4. It is likely that the correlation results could have been greater if collocations with a larger range of frequency had been included. That said, the target collocations of this study were selected because they are high-frequency and thus useful to learners.

It should be added that the frequency effect on recognition knowledge of academic collocations in the present study should be interpreted with caution. As presented in Chapter 6, the lack of correlation was consistent in the case of the AC Recall Test, irrespective of with or without five items with a frequency greater than 1,000. Meanwhile, the result for the AC Recognition Test was changed from a modest correlation ($r_s = .34$, $p < 0.01$) to a non-significant correlation ($r_s < .18$, $p > 0.1$) after removing those five high-frequency items. It was possible that the ceiling effect found in the AC Recognition Test (see Chapter 5) had an impact on the correlation result. In other words, because the AC Recognition Test was generally very easy for the participants of this study, relationships between frequency and recognition knowledge of academic collocation could not be confirmed. Overall, the current study proposes that the frequency effect on knowledge of academic collocations is not as straightforward as in the case of single words (see Schmitt et al., 2001; Webb & Chang, 2012).

7.1.5 Non-significant relationship between time spent in an ESL context and knowledge of academic collocations

The final finding of the present study related to the development of academic collocation knowledge is that the time spent studying in an English-speaking context has a non-significant relationship with this knowledge. One possible explanation is that the length of immersion might not be long enough to improve learners' knowledge of academic collocations. At the time of testing, most of the ESL participants in this study had spent three to four years at university in New Zealand. Li and Schmitt's (2010) longitudinal study shows that the development of second language (L2) learners' collocation knowledge tends to be slow, even for highly proficient students. Laufer (2011) also found that the use of collocations is problematic for L2 learners, irrespective of the years of instruction they received in L2.

Another possible reason is that the ESL context consists of many variables that can affect language development, including first language (L1) backgrounds and the amount of interaction using L2. Research has found that the degree of congruency between L1 and L2 has an impact on the acquisition of collocations (e.g., Cao & Badger, 2021; Nesselhauf, 2005; Nguyen & Webb, 2017). If the two languages are similar or congruent, learners may acquire L2 collocations more easily. Furthermore, the amount of L2 interaction is also important for the development of academic collocation knowledge. According to Adolphs and Durow (2004), learners need to have an extensive exposure to the L2 environment through social and cultural activities, as well as frequent contact with native speakers to improve knowledge of formulaic sequences. Therefore, learners who study in an ESL context but do not actively engage in L2 use might not have much improvement in collocation knowledge. Taken together, the ESL context displays great variability, which may contribute to the lack of correlation with learners' academic collocation knowledge in this study.

So far, this section has indicated that the development of knowledge of academic collocations does not follow the same pattern as that of general collocations. Therefore, a separate measure for academic collocation knowledge, such as the ACTs in this study, is a must. The discussion based on the ACT results in this study has contributed to our understanding of the development of academic collocation knowledge and factors that affect learning. The following section discusses the ACTs more directly by looking into their creation through the use of a published word list.

7.2 What are the opportunities and challenges of developing the ACTs from a published word list?

The ACTs were developed from the Academic Collocation List (ACL) (Ackermann & Chen, 2013) based on results of the evaluation process to compare two lists of academic collocations (see Chapter 4, Section 4.2). Using a corpus-based word list to develop the ACTs has both benefits and drawbacks. Test-developers need to be well aware of these aspects for future application of this wordlist-based approach in testing multiword units.

7.2.1 Opportunities of developing the ACTs from a corpus-based word list

Developing the ACTs using the ACL (Ackermann & Chen, 2013) had four major advantages. The first benefit involves the representativeness of test items. As described in Chapter 4 (Section 4.2.1), the ACL (Ackermann & Chen, 2013) employed both objective and subjective methods (i.e., computer extraction and human intuition) to select items that are representative of the academic corpus and pedagogically relevant. The development of the ACL (Ackermann & Chen, 2013) reflects the interaction among three important dimensions in corpus research, as mentioned by Martinez (2019): the sample, the application and the researcher(s). That is, in producing the ACL, Ackermann and Chen (2013) (i.e., the researcher dimension) interpreted the corpus data (i.e., the sample dimension) to fit its intended pedagogical purpose (i.e., the application dimension). This compilation process distinguishes the ACL from other lists of academic collocations that rely purely on a quantitative approach, such as the study of Lei and Liu (2018) which seems to lack the interaction between the researcher and application dimensions. The wordlist evaluation process in Chapter 4 (Section 4.2.3) also revealed that the ACL achieved stable coverage over different academic corpora (i.e., the academic corpus used to develop the list and the independent corpus used in the present study). Taken together, selecting test items for the ACTs from such a well-made word list can achieve high representativeness of frequent and pedagogically relevant academic collocations.

The second advantage of wordlist-based test development is related to the interpretation of test scores. A test will only be useful when test users can understand the meaning of the test results (i.e., what the test scores can reveal about a learner's knowledge). Because this study based the ACTs on the ACL (Ackermann & Chen, 2013), the test scores can be interpreted as to what extent test-takers know high-frequency academic collocations in this list. This is different from previous studies such

as Voss (2012) whose items were sampled directly from a corpus and the scores were vaguely interpreted as an indication of learners' academic collocation knowledge. Because there are a vast number of collocations, a limited number of test items such as 35 items in Voss (2012), cannot represent collocations in a whole corpus, which makes the generalisation of the test scores problematic. The interpretation of the scores on the ACTs is more meaningful in a way that it is based on a finite list of academic collocations rather than a huge number of items in a corpus.

The third advantage of using a corpus-based word list for test development is that it closes the gap between the fields of wordlist development and vocabulary testing. The present study is the first that applied a published list of academic collocations in testing. A well-made word list such as the ACL (Ackermann & Chen, 2013) requires a lot of time and effort. Even with the use of a computer to automatically extract items based on a set of pre-specified criteria, the list still involves a great deal of manual checking. The results of such a wordlist development study will therefore provide reliable sources for the vocabulary testing field. This direction of using word lists in vocabulary assessment however has mostly been applied with frequency-based lists of individual words with the creation of the Vocabulary Levels Test (Schmitt et al., 2001), the Vocabulary Size Test (Nation & Beglar, 2007), the Academic Vocabulary Test (Pecorari et al., 2019), just to name a few. The creation of the ACTs is a response to Nation (2016) and Gyllstad and Schmitt's (2018) suggestion of using word lists for test development.

Last but not least, using a word list with a pedagogical purpose such as the ACL (Ackermann & Chen, 2013) to develop the ACTs creates a strong connection between vocabulary instruction and assessment. This has been seen in the case of the Academic Word List (Coxhead, 2000) and the Academic Vocabulary List (Gardner & Davies, 2014) which have a significant impact on EAP teaching and testing. When results of a test show that learners need more support in developing their academic vocabulary knowledge, the list that was used to develop the test of academic words will become a clear goal for learners to work on. The same principle could also be applied to the ACTs and the ACL (Ackermann & Chen, 2013). In addition to the list itself, teachers and learners can benefit from the resources that have been developed based on the ACL (see Chapter 8, Section 8.2.1.2). This connection between testing and teaching gives more credit to the practicality of the ACTs.

The advantages outlined in this section which arose from the development of the ACTs based on the ACL (Ackermann & Chen, 2013) are encouraging. Similar applications of lists of multiword units to vocabulary assessment would also bring benefits not only to the test itself in terms of item representativeness and score interpretation, but also to the wider connections between the fields of wordlist studies and vocabulary testing, as well as between vocabulary instruction and assessment.

7.2.2 Challenges of developing the ACTs from a corpus-based word list

The development of the ACTs from the published list of Ackermann and Chen (2013) also involved some obstacles to be overcome. The first challenge came from the process of evaluating two academic collocation lists to identify the better source of items for the test development. One of the difficulties in this process was the lack of an applicable framework for lists of multiword units. Nation's (2016) wordlist evaluation framework was applied in the current research, although it was designed specifically for lists of individual words. This study adapted Nation's (2016) framework to be used with lists of multiword units.

The next disadvantage of basing the ACTs on a word list was that neither collocation lists was ready to be used for the test development due to the lack of frequency information. The AECL (Lei & Liu, 2018) is not publicly available, and only a small sample of the list is provided in the published article. The authors kindly shared the AECL to be used in this study, although the frequency information was not included in the list. Similarly, the ACL (Ackermann & Chen, 2013) was published without the frequency information. Frequency is an important factor for vocabulary acquisition (Coxhead, 2019; Dang & Webb, 2016; Vilkaitė-Lozdiene & Schmitt, 2019; Webb & Nation, 2017). Therefore, frequency was also taken into consideration when developing the ACTs to investigate its effect on learners' knowledge. When a list is divided into different frequency bands, a random selection of items from each frequency band will ensure the items selected are not biased towards any specific band, but represent a variety of frequency bands (Nation, 2016). As a result, a word list published with frequency information will support test development with frequency-based item selection, as in the case of the phrase test developed from the Phrasal Expressions List (Martinez, 2011) and the phrasal verb tests (Garnier & Schmitt, 2016; Sonbul et al., 2020) based on the Phrasal Verb Pedagogical List (PHaVE List) (Garnier & Schmitt, 2015). Without the available frequency information of academic collocations in the ACL (Ackermann & Chen, 2013), I had to manually

search for this information in an independent corpus for each item in the list. This was a major obstacle but a vital step in the research. It was not a small task given the list size of 2,469 items.

The final challenge stemmed from the fact that limitations of the ACL (Ackermann & Chen, 2013) are inherent in the limitations of the ACTs developed from the list (see more in Chapter 8, Section 8.3). First, the ACL includes lexical collocations only so the ACTs do not cover grammatical collocations (e.g., *lack of* or *focus on*) that can also be useful for pedagogy. Another criticism for the ACL (Ackermann & Chen, 2013) is that the list made extensive use of human intuition and might have removed some useful collocations. This critique has been thoroughly discussed in Chapter 4 (Section 4.2.1). Having said that, not all the criticisms are necessarily negative. The methodology which included the subjective judgement in the development of the ACL (Ackermann & Chen, 2013) matched the testing purpose in the current study, which was to only include items that are highly frequent and pedagogically relevant.

This section has highlighted some possible obstacles in evaluating multiword-unit lists, the lack of frequency information of the published lists and potential criticisms associated with the source list that future studies may encounter. Being aware of these obstacles is important for the decision to base a test on a word list.

7.3 How can the test development and validation frameworks support the creation of the ACTs?

This research is among very few studies that apply the evidence-centred design (ECD) framework to developing and the argument-based framework to validating vocabulary tests. This section first discusses how the ACTs are different from the other collocation tests as a result of using these frameworks. The section then delves deeper into the validity argument of the ACTs and discusses it in terms of the most important inferences and how necessary they are to be fully supported.

7.3.1 How the ACTs are different from other collocation tests?

With the application of the ECD framework, the ACTs could overcome some of the issues of previous tests related to selecting collocation items and designing test tasks that could appropriately capture the target knowledge. First, the domain analysis is the key layer of the ECD that contributed

to the robustness of the item selection process of the ACTs. Within this layer, corpus-based word lists of academic collocations were identified and evaluated to select the best source of items for the ACTs (see Chapter 4, Section 4.2). The domain analysis is essential to ensure the collocation items are representative of those in the written academic domain. Despite its importance, many previous studies on designing collocation tests did not pay enough attention to this step and did not clarify the domain from which the collocations were taken (e.g., Bonk, 2000; Gaballa & Al-Khayri, 2014; Keshavarz & Salimi, 2007). The item selection based on the domain analysis, therefore, helps the ACTs to stand out from the available tests on collocations.

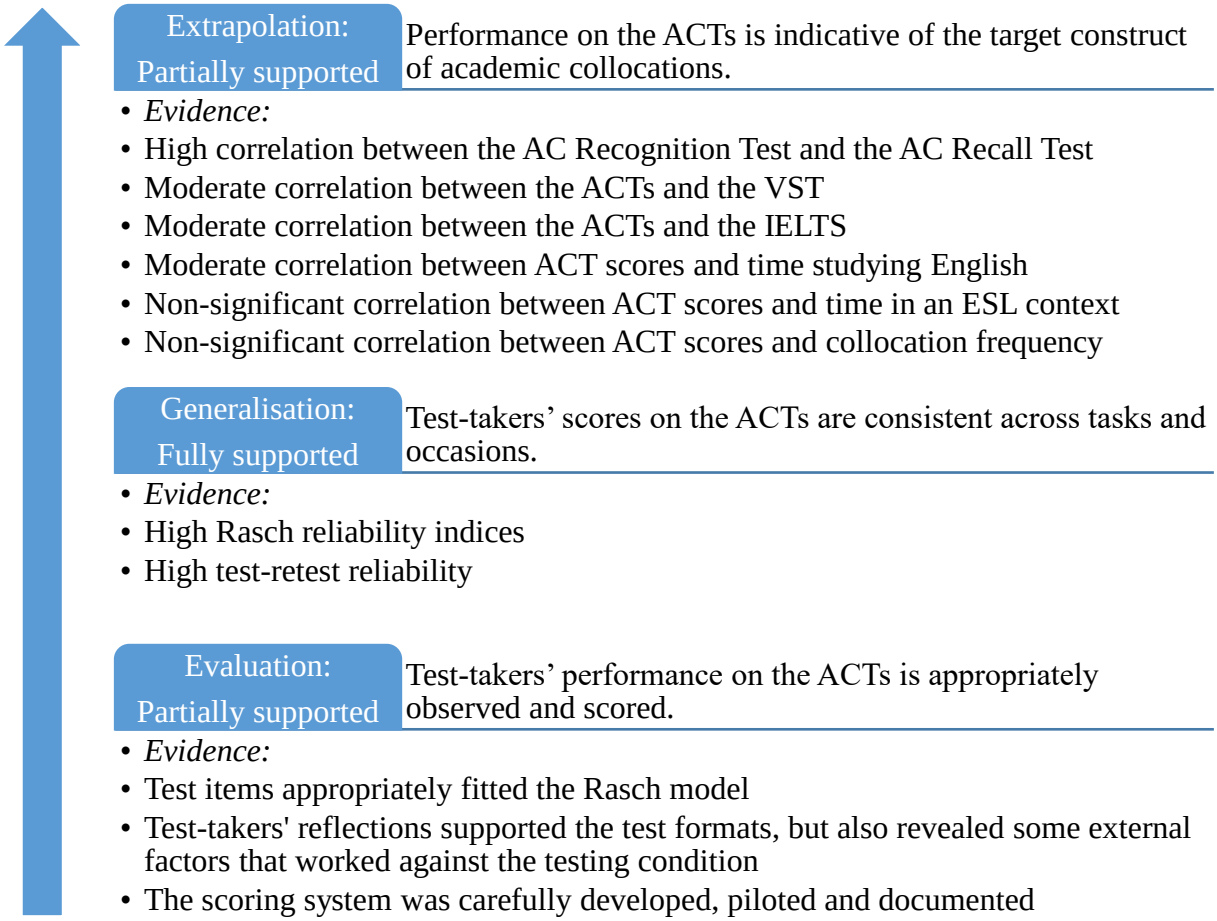
Employing the ECD also allowed me, as a test developer, to have a clear view of what is being measured through the test task and what knowledge is required to answer the test items in the ACTs, as specified in the domain modelling layer. Although the multiple-choice format used for the AC Recognition Test and the gap-filling format used for the AC Recall Test were not new, they were adopted to measure the target knowledge of academic collocations more precisely. While previous measures that employed these formats, such as those in Voss (2012) and Wongkhan and Thienthong (2020), broke down collocations and tested only one word of each collocation pair (i.e., test-takers were asked to choose or fill in just one word instead of an entire collocation), the test items in the ACTs always present the collocations as holistic units. This is important because the collocation construct was not mismatched with another construct such as knowledge of individual words. Overall, the design of the ACTs was strengthened because the ECD framework provided the foundation to ensure that the target knowledge of academic collocations could be assessed through specific tasks employed in the ACTs.

Last but not least, the validation process of the ACTs following the argument-based framework adds to the rigorousness of the tests. Previous tests have been released without validation evidence (e.g., Bahns & Eldaw, 1993; Fernández & Schmitt, 2015; Nguyen & Webb, 2017; Wongkhan & Thienthong, 2020), but the ACTs have undergone a validation process that provides future test users with both quantitative and qualitative evidence to evaluate the test quality. All the research findings are connected in a comprehensive framework to assess the validity of the inferences made about the ACTs (see Figure 7.1). The three inferences included in the validity argument for the ACTs are Evaluation, Generalisation and Extrapolation. The Evaluation inference was based on the

characteristics of the test items, the testing conditions and the scoring procedures. The Generalisation inference relied on the consistency of item measures and testing occasions. The Extrapolation inference counted on the correlations with other measures, with the frequency of academic collocations and with learning experience. The inferences and the collected evidence are summarised in Figure 7.1. Overall, all the three inferences were supported to a certain extent.

Figure 7.1

Validity Argument for the ACTs With Main Findings



The argument-based validation employed for the ACTs is similar to that used by Voss (2012) and Chen (2019). The application and presentation of this framework, however, are simplified so that the study results are highlighted to provide a better understanding of the development of academic collocation knowledge, and at the same time, ensure the interpretations of the test validity. The

following section will look into this issue more closely with a discussion of the most important inferences in a validation framework.

7.3.2 What are the most important inferences in the validation framework of the ACTs?

The present study suggests that Evaluation and Generalisation inferences are most needed in building the validity argument for the ACTs in particular, and for a vocabulary test in general. The Evaluation inference is concerned with the core elements of a test, including test characteristics, testing condition, and scoring system. The Generalisation inference is related to reliability which is an important assessment of test quality. As illustrated in Figures 7.1 and 7.2, a validity argument is a sequence of inferences from a lower-level to a higher-level one. If the former is not supported, the latter cannot stand. For example, if it had been found that the test characteristics of the ACTs were not appropriate to measure academic collocation knowledge (Evaluation inference), or the tests were not reliable (Generalisation inference), it would be pointless to further investigate whether the test scores correlated with other factors (Extrapolation inference), or if the test scores could be used for decision making (Utilisation inference). The Evaluation and Generalisation inferences, therefore, lay the foundation for higher-level inferences such as Extrapolation and Utilisation inferences.

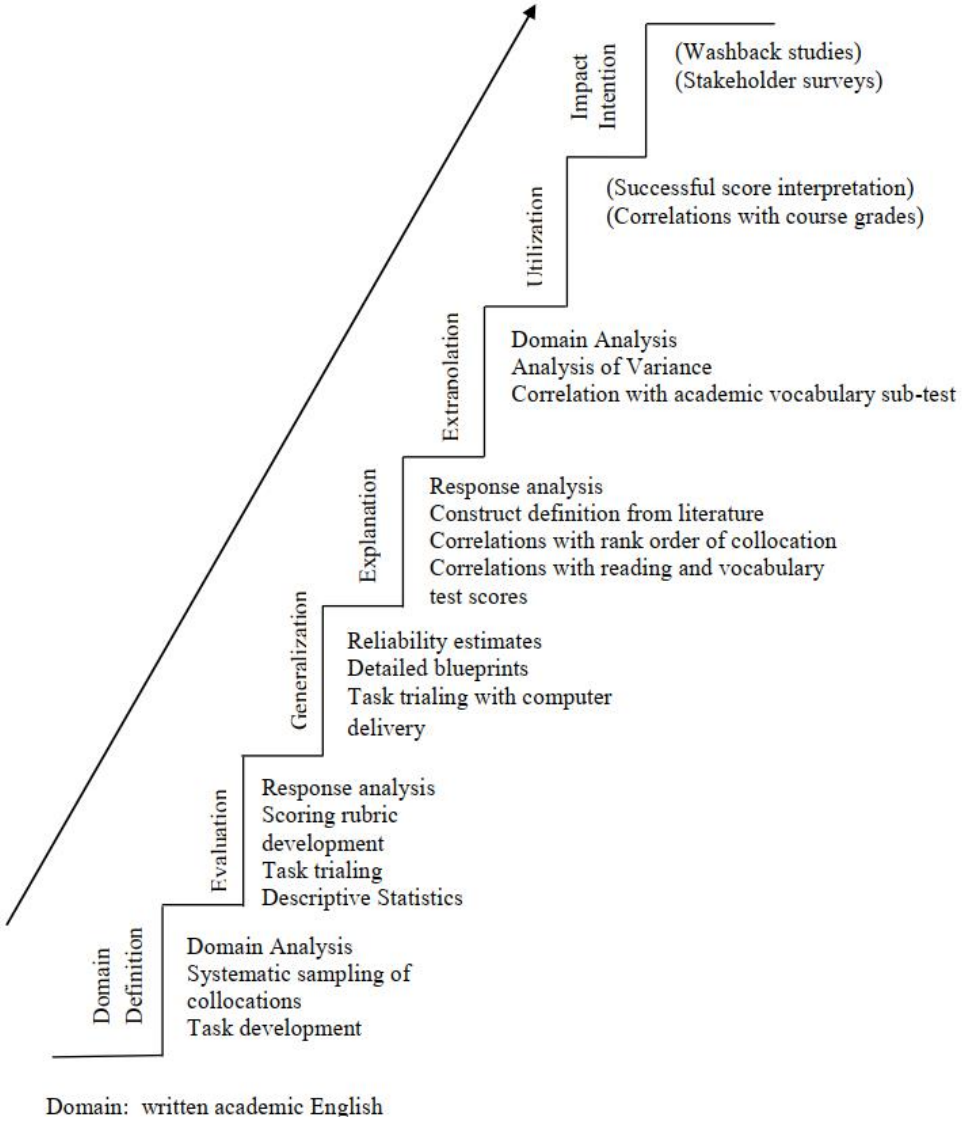
There are several reasons why the present research comprised only three inferences (Figure 7.1), instead of seven as in Voss (2012) (see Figure 7.2) or six as in Chen (2019) (without the Impact intention). First, as this was the initial validation of newly created tests, it was more practical to start with the most fundamental inferences. As previously explained, the Evaluation and Generalisation inferences are the two most needed elements in any validity argument of a vocabulary test. The Extrapolation inference was added to the validation framework of the ACTs to provide a better understanding of the relationship between academic collocation knowledge and other factors such as vocabulary size, general language proficiency and English learning experience.

Second, although the other inferences were not included in the validity argument of the ACTs, they had been considered in some ways. The Domain definition inference did not appear in the validation process because it overlapped with the domain analysis and domain modelling layers of the test development framework. Chapter 4 in this thesis has reported the domain analysis, the systematic sampling of academic collocations and the task development of the ACTs, which constituted the

Domain definition inference in Voss's (2012) validation framework (Figure 7.2). The Explanation inference was absent from the validity argument of the ACTs because the present study adopted Kane's (2004) framework which does not distinguish between Explanation and Extrapolation inferences, but all correlation analyses are related to Extrapolation (see Chapter 2, Section 2.5.2). The Utilisation and Impact intention inferences in Figure 7.2 are only tentative because Voss (2012) did not conduct empirical research to obtain evidence for them. Similarly, only when the ACTs are being employed in a course or a program can inferences be made about their use and impact.

Figure 7.2

Validity Argument for the Collocation Ability Test With Collected Evidence (Voss, 2012, p.171)



Third, in an attempt to promote the argument-based validation framework for further application to other vocabulary tests, this study deliberately simplified the framework to the most basic inferences. One possible reason why this framework is now the mainstream in language testing but still not a common practice in vocabulary assessment is because of its complexity with many layers of concepts. Although Chen (2019) is a large project which fully applied the validation framework with six inferences and a diverse set of evidence, this unpublished doctoral dissertation is particularly not easy to read. Perhaps with a great deal of evidence to connect with a complicated network of inferences, Chen (2019) struggled to summarise the research findings and present them in a coherent framework as the present study (Figure 7.1) and Voss (2012) (Figure 7.2) did. Interpreting 878 pages of Chen's (2019) thesis to understand the framework and recognise its value is a real challenge. As a result, the present study aimed to keep the validation framework as brief and straightforward as possible in order to motivate other vocabulary researchers and practitioners, especially those attempting to apply this framework for the first time (such as me).

7.3.3 What if an inference in the validation framework of the ACTs was not fully supported?

Ideally, all the inferences in a validation framework should be fully supported, although this is not always the case. As shown in the validity argument of the ACTs in Figure 7.1, the Evaluation and Extrapolation inferences received partial support. The purpose of this section is to discuss why they were not fully supported, what the impacts were, and how their level of support can be improved.

Let us look at the Evaluation inference first. There were two reasons why this inference was not completely justified. First, the test delivery switched from direct administration to online testing due to the outbreak of the COVID-19. Online delivery had some disadvantages. The post-test interviews revealed some interruptions during the test-taking time and dictionaries use by some test-takers. Second, as the two academic collocation tests targeted the same 60 items embedded in the same context sentences, a possible priming effect could not be ruled out. Test-takers may have provided a correct response to the recognition test based on their memory from the recall test instead of their own knowledge. Remembering the initial letter hint in the AC Recall Test may have helped test-takers to choose the correct option in the AC Recognition Test which shared the same initial letters.

The above-mentioned factors might cause a threat to the validity of the test results and weaken the Evaluation inference. If these were high-stakes tests and the test results were used for making a decision, the limitations of the testing condition would have a negative impact on test-takers. For example, they might be assigned to the wrong course level. That said, the ACTs were low-stakes tests for diagnostic purposes, the influence was not that tremendous. For the participants in this study, the results of the ACTs did not affect them in any way. Moreover, constraints associated with the test administration did not directly reflect the quality of the tests themselves. If it had been an issue with the test characteristics or the scoring system, the Evaluation inference would not have been held eventually.

To improve the level of support for the Evaluation inference, one way would be to explicitly state in the test instructions that test-takers are encouraged to complete the tests in a place without any noise disturbance or distractions, and they are not allowed to consult dictionaries or any other external resources. A better way still would be administering the tests under the supervision of researchers or assistants. Furthermore, it is possible that spacing the two academic collocation tests out by a few weeks or months would lessen the priming effect.

Turning now to the Extrapolation inference, it was not fully warranted because knowledge of academic collocations was found to have non-significant relationships with the frequency of academic collocations and with time spent in an L2 learning context. This was the result of wrong assumptions about the indication of academic collocation construct. It was presumed that the higher the frequency of academic collocations, the easier the test items would be, and more time spent in an ESL context would result in better academic collocation knowledge. However, as previously discussed in Section 7.1, these assumptions were not true because of the frequency nature of the academic collocations and the high variability of the ESL context.

Although not all of the backings were found for the Extrapolation inference, there was no serious impact on the results of this study. First, the most important warrant of this inference was supported in that, scores on the ACTs were correlated with scores on other tests of similar constructs, such as the VST (Nation & Beglar, 2007) and the IELTS (Section 7.1). This is critical because it clearly demonstrated that the construct of academic collocations is related to vocabulary size and general language proficiency. Second, the non-significant correlations found in this study also contribute to

understanding the nature of the academic collocation knowledge which is not as straightforward as that of single words.

To improve the level of support for the Extrapolation inference, further correlation analyses could be conducted. One possible option would be to link the scores on the ACTs with learners' reading and writing skills to examine the relationship between learners' knowledge of academic collocations and their ability to comprehend and produce academic texts. This would provide evidence to assess whether scores on the ACTs could be used to predict the ability to apply this knowledge in a real target language use domain. Furthermore, other tests of similar constructs could be used for correlation analyses instead of the VST (Nation & Beglar, 2007), such as the Phrasal Vocabulary Size Test (Martinez, 2011) or tests of general collocations (e.g., Gyllstad, 2009; Nguyen & Webb, 2017). These are all tests on multiword units, which means the relationships between them would be expected to be stronger.

7.4 Chapter summary

This chapter discussed three major themes drawn from the findings of the thesis. The first theme – the development of academic collocation knowledge – is a burning issue in vocabulary studies. The other two themes – basing a test on a word list and applying test development and validation frameworks – are underexplored areas of vocabulary research. The next chapter concludes the thesis by looking at the contributions of the present study, the implications for research and pedagogy, the limitations and directions for future research.

Chapter 8 Conclusion

The present study has addressed a major gap in applied linguistics in the area of written academic collocations and contributed to research on this group of lexis. This study set out to design two tests to examine recognition and recall knowledge of academic collocations in two educational contexts (Vietnam and New Zealand). Following the evidence-centred design framework (Mislevy & Yin, 2013), the two Academic Collocation Tests (ACTs) were developed from the Academic Collocation List (ACL) (Ackermann & Chen, 2013) based on the results of a wordlist evaluation process. After that, the validation process embedded within the argument-based approach (Kane, 2004, 2013) involved the assessment of three main inferences made about the ACTs (i.e., Evaluation, Generalisation and Extrapolation). This chapter begins with the contributions of this study in different areas: vocabulary studies, vocabulary testing and wordlist research in Section 8.1. This is followed by implications for different stakeholders including EAP teachers, test developers, EAP material designers and list developers in Sections 8.2. The limitations and directions for future research are then addressed in Sections 8.3 and 8.4. The chapter ends with reflections on my PhD journey in Section 8.5.

8.1 Contributions

This study has combined several areas of research: vocabulary studies, vocabulary testing and wordlist research. This multidimensional study, therefore, contributes to each area in different ways. In the following sections, I will discuss these contributions for each area in turn.

8.1.1 Contributions to vocabulary studies

The present research has made two valuable contributions to vocabulary studies. Its biggest contribution is the Academic Collocation Tests (ACTs) which are ready-to-use measures to support EAP pedagogy and potentially be used in future research. These two instruments can assess recognition and recall knowledge of academic collocations. EAP teachers can use the tests for diagnostic purposes, that is, to estimate learners' current knowledge of academic collocations and use the findings to inform their curriculum. In addition to this important pedagogical contribution to the field, the ACTs are useful research tools which can be used to investigate how knowledge of academic collocations correlates with factors such as overall vocabulary size or general language

proficiency. The tests can open up opportunities for future studies to further investigate the nature of academic collocation acquisition.

Another major contribution of the present study is that it enhances our understanding of the development of academic collocation knowledge in different contexts and in several ways. The EFL/ESL students in this study demonstrated substantial recognition knowledge but less recall knowledge of academic collocations. This research has shown that although recall knowledge lagged behind recognition knowledge, these two aspects of knowledge were strongly correlated. The current findings have also pointed out that knowledge of academic collocations moderately correlated with overall vocabulary size, general language proficiency and years of English study. However, the relationships were negligible when academic collocation knowledge was correlated with the frequency of collocations and years in an ESL context. This study contributes to an overall picture of the development of academic collocation knowledge and factors that affect learning.

8.1.2 Contributions to vocabulary testing

The current research has two significant contributions to vocabulary testing by basing tests on a word list and utilising a validation framework. First, this study has been one of the first to develop tests of academic collocations from a corpus-based word list. The development of the ACTs based on the ACL (Ackermann & Chen, 2013) provides encouragement for a similar application in vocabulary testing. There are benefits to this approach, including item representativeness, score interpretation, and connections between the fields of wordlist studies and vocabulary testing, as well as between vocabulary instruction and assessment (see Chapter 7, Section 7.2). This study introduces six test development steps: 1) Identify available word lists of academic collocations, 2) Evaluate the academic collocation lists, 3) Sample the academic collocation items from the selected list, 4) Select the test formats, 5) Write the test items, and 6) Pilot and finalise the tests. These steps were embedded in the evidence-centred design framework (Mislevy & Yin, 2013). Details of each step have been described in Chapter 4. The method used for the test development in this study lays the groundwork for the further application of multiword-unit lists into developing tests.

Second, the current study has advanced the field of vocabulary testing by adopting the argument-based validation framework (Kane, 2004, 2013) to provide an in-depth assessment of the ACTs. The

tests have undergone a validation process that presented both quantitative and qualitative evidence to evaluate three major inferences in the validation framework. The Evaluation inference focused on the characteristics of the ACTs which were appropriate for test-takers to demonstrate their knowledge of academic collocations. The examination of the Generalisation inference revealed that the ACTs were highly reliable. The investigation of the Extrapolation inference helped to discover factors that affect the acquisition of academic collocation knowledge. The present study has provided support for the value of the argument-based validation framework and generated motivation for test developers to apply it to other vocabulary tests.

8.1.3 Contributions to wordlist research

The present study has two important contributions to wordlist research: an evaluation framework for multiword unit lists and an indication of the significance of academic collocation lists. This study appears to be the first to compare and evaluate two lists of academic collocations. This thesis has provided a model of evaluating lists of multiword units and builds the foundation for future research into the field of wordlist research. Prior to this study, there was no clear path for researchers to evaluate different lists of multiword units. This study suggests an approach for a thorough assessment of the lists by combining different methods of evaluation, including Nation's (2016) adapted framework, lexical constituents comparison and lexical coverage analysis. The evaluation framework adapted from Nation (2016) first provides an overall picture of how the lists are developed and validated. The lexical constituents comparison gives a closer look into how the words on the lists are similar and different. The lexical coverage analysis then helps to point out the group of most frequent items which should be the initial target of learning. By modelling the practice of evaluating word lists, this study highlights the importance of this work and provides a ready-made framework to guide or inform similar attempts in wordlist development studies.

The final contribution of this study is to enhance our understanding of the value of the ACL (Ackermann & Chen, 2013) and the Academic English Collocation List (AECL) (Lei & Liu, 2018). Granger and Larsson (2021) express concern that academic multiword unit lists have not had a great impact on EAP teaching and assessment when compared to single-word academic lists. The evaluation of the two academic collocation lists in this thesis is hoped to contribute to the wider application of the lists in the future. These are valuable resources to provide a basis for classroom

instructors, EAP textbook designers and test developers to select the list that best suits their purposes. Those stakeholders may wish to differentiate between the two lists but may not have time and resources to conduct the evaluation on their own. For example, to compare the list coverage, access to a large corpus and an available analysis tool is a must. Additionally, list evaluators also need to know how to run the analysis and interpret the results. This specialised knowledge and skills take time to develop. Therefore, this study provides research-based evidence for end-users of the lists to base their decisions on the advantages and disadvantages of using different word lists.

8.2 Implications

The findings from the present study have led to a number of practical implications for EAP teachers, material designers, test developers and wordlist developers.

8.2.1 Implications for EAP teachers

This study provides useful pedagogical implications for language teachers on how to use the ACTs in EAP contexts and how to develop learners' academic collocation knowledge based on the test scores. This section looks at each of these areas in turn.

8.2.1.1 Using the ACTs in EAP courses

The ACTs can be used in several ways to support EAP. First, the ACTs can be used as one single administration at the beginning of an English course to help teachers plan an appropriate course of instruction. These tests are beneficial for learners who plan to study at an English-medium university as a way to diagnose their level of knowledge of academic collocations at an early stage so that proper support could be provided. The tests are also suitable for learners who are already studying in English-medium programs but having difficulty with academic reading and writing.

Second, the ACTs may be used as pre- and post-tests to measure learning gains. For example, the tests could be administered before a language course and then again after ten or more weeks to check if learners' levels of academic collocation knowledge improve. Based on the test results, teachers may adjust the focus on academic collocations as one of the components in the teaching syllabus. More attention to this vocabulary might be needed if the learning gain is found to be minimal and slow.

Third, depending on the focus and the time availability of a course, teachers can use one or both of the ACTs. Each test has its own value in indicating different aspects of academic collocation knowledge. The AC Recognition Test can be used to investigate learners' ability to recognise academic collocations in a given context, and the AC Recall Test can be employed to examine learners' capacity to produce academic collocations to provided prompts. Due to the strong correlation between the two tests (see Chapter 6, Section 6.2.1), the AC Recognition Test could be used as a more practical instrument than the AC Recall Test in terms of administration and scoring to provide an instant capture of learners' academic collocation knowledge.

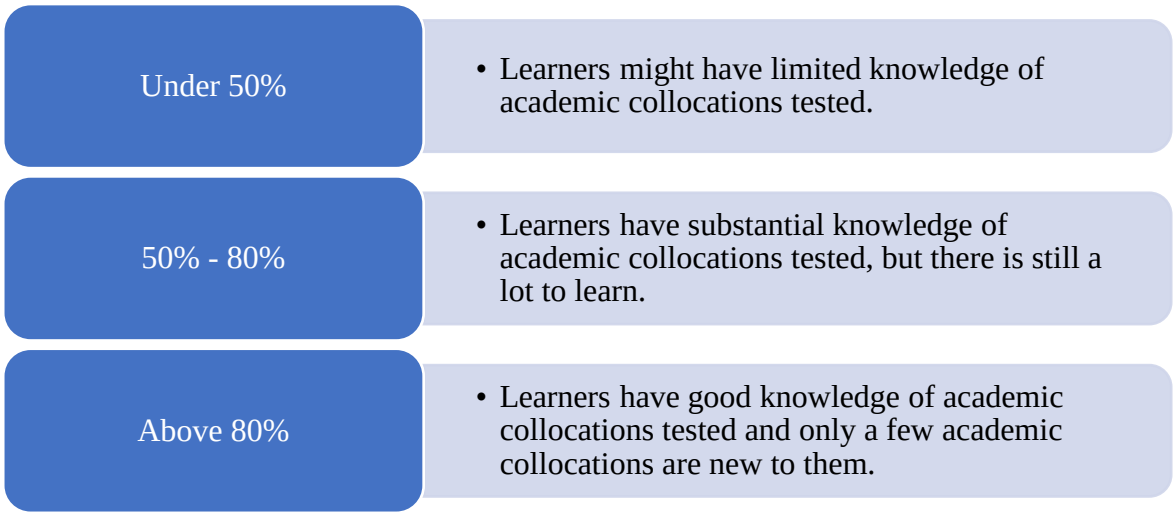
8.2.1.2 Developing learners' academic collocation knowledge based on the ACT scores

One way to decide on whether learners need support with developing their academic collocation knowledge is to look closely at the ACT scores and apply a cut-off score. Taking the group of participants in this study as an example, the results in Chapter 5 showed that, on average, the students scored almost 75% on the AC Recognition Test and 43% on the AC Recall Test. If 80% was selected as the cut-off score for passing the tests, these students would definitely need further support in developing their recognition and recall knowledge of academic collocations. The exact cut-off score should be determined by teachers to match their testing purposes. For example, the cut-off score for checking whether learners can understand collocations in EAP textbooks could be lower than that for assessing learners' ability to comprehend collocations in authentic academic texts such as research articles. This is because academic collocations may appear less frequently in EAP textbooks (Coxhead et al., 2020) than in scholarly papers.

This thesis proposes cut-off scores (see Figure 8.1) for general diagnostic purposes in order to estimate the extent to which learners need support with their academic collocation knowledge. The 80% threshold follows the suggestion for previous vocabulary tests, including the Vocabulary Levels Test (Schmitt et al., 2001), while the 50% cut-off marks the point of below or above average. It should be noted that further research is needed to provide theoretical or empirical evidence to justify these cut-off scores.

Figure 8.1

Recommendations for Interpretations of ACT Scores



Based on the test scores, teachers can give students some practical guidelines on how to improve their knowledge of academic collocations. Table 8.1 gives suggestions on activities which aim to increase learners' noticing of academic collocations and encourage them to use academic collocations in their writing at university. For the lowest level (i.e., scores under 50%), teachers can start with introducing the ACL (Ackermann & Chen, 2013) to learners and discuss the concept of academic collocations. The activities suggested in Table 8.1 for this level mainly focus on raising learners' awareness about words that make up a collocation as a holistic unit in academic texts. For the higher level (i.e., scores between 50% and 80%), the suggested activities aim to create opportunities for learners to develop the routines of learning new academic collocations from readings and using them in academic writing. For the highest level (i.e., scores above 80%), learners are encouraged to maintain the routines of learning and using academic collocations.

Even though independent learning of academic collocations is important, the role of teachers is still critical in modelling the activities in the classroom (e.g., how to use the online resources) and supervising learning outside the classroom (e.g., asking learners to report newly learned collocations or submit writing portfolios). Online resources have been found to be effective in helping learners extract and look up multiword units, which in turn improves their acquisition (see Bui et al., 2020).

These tools such as the ACL highlighter and the COCA corpus tool (Table 8.1) can also be employed for developing knowledge of academic collocations.

Table 8.1

Suggestions for Teachers and Learners for Improving Academic Collocation Knowledge Based on Test Scores

ACT scores	What could learners do?	
	The AC Recognition Test	The AC Recall Test
Under 50%	• Become familiar with the idea of academic collocations and top 500 items in the ACL (Appendix H); • Identify academic collocations with the aid of online resources.	• Learn how academic collocations are used in academic writing from concordance lines in the COCA corpus tool[1]; • Practise writing sentences with academic collocations in the ACL.
50% - 80%	• Continue to work with the ACL using ACL highlighter[2]; • Find more academic collocations using the COCA corpus tool.	• Develop a routine of using academic collocations in writing assignments or academic texts; • Analyse your own texts with the ACL highlighter regarding the quantity and variety of academic collocations used.
Above 80%	• Maintain exposure to academic texts; • Continue to identify and learn new collocations from academic texts.	• Maintain the routine of using academic collocations in academic writing; • Try to increase the number and the variety of academic collocations used in writing.

Note. The online resources can be found at the following websites:

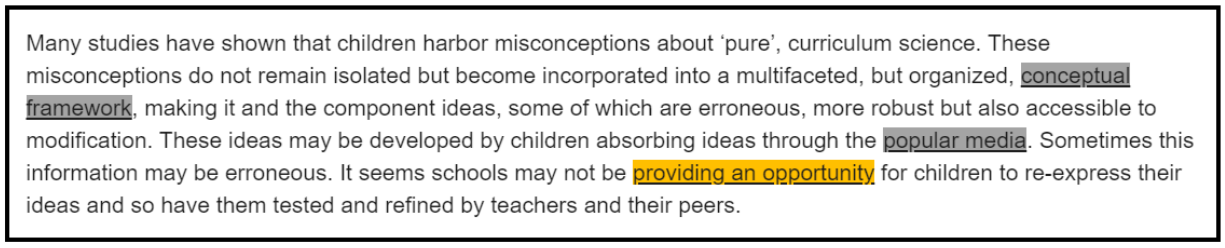
[1]The COCA corpus tool: https://www.english-corpora.org/coca/

[2]The ACL highlighter: https://www.eapfoundation.com/vocab/academic/acl/highlighter/

With the ACL highlighter, users can enter a text and identify all the ACL items that appear in the text (see Figure 8.2 for an example). This tool will be beneficial for various learners at different cut-off levels (see Table 8.1). They can discover more collocations from analysing academic texts with the ACL highlighter. In this way, collocations can be recognised and learned in a meaningful context. The ACL highlighter also helps learners to analyse their own writing and keep track on the academic collocations that they produce. A careful record can show learners whether they can increase the number of academic collocations being used in a single piece of writing and whether they employ a repeated number of collocations or a wide range. Such tracking may raise learners' awareness of their writing patterns and motivate them to learn more academic collocations.

Figure 8.2

Example of an ACL Highlighter Output Using an Extract in Cambridge IELTS 4: Examination Papers From University of Cambridge ESOL Examinations (2005)



Similarly, the COCA corpus tool can be employed in several ways to benefit learners' recognition and recall knowledge of academic collocations (see Table 8.1). The *Collocates* function of this tool allows users to fill in one word of a collocation pair (the node word) and all possible collocates of that node word will be displayed. Users can refine the results by setting some criteria such as using words found in a specific corpus (e.g., Academic, Fiction, Spoken, etc.), choosing the span of the search or parts of speech of the collocates and setting the minimum frequency of the collocations. Figure 8.3 shows example results of a search for noun collocates of the node word "conceptual" in COCA Academic corpus with the span of one word to the right of the node word. In addition to identifying collocations, the COCA corpus tool also provides concordance lines which give the context in which the collocations are used. For example, Figure 8.4 shows how the academic

collocation “conceptual framework” is used in real academic contexts through the concordances in COCA.

Figure 8.3

Top Ten Collocates of the Node Word “Conceptual” Found in COCA Academic

HELP			FREQ		ALL	%	MI	
1	☐	FRAMEWORK	804		14855	5.41	10.06	
2	☐	MODEL	402		59078	0.68	7.07	
3	☐	UNDERSTANDING	239		36434	0.66	7.02	
4	☐	FRAMEWORKS	105		1842	5.70	10.14	
5	☐	CHANGE	101		53814	0.19	5.21	
6	☐	KNOWLEDGE	90		49966	0.18	5.15	
7	☐	ART	87		38047	0.23	5.50	
8	☐	MODELING	79		5933	1.33	8.04	
9	☐	MODELS	78		24343	0.32	5.98	
10	☐	ISSUES	66		39831	0.17	5.03	

Figure 8.4

Example of Concordance Lines of the Collocation “Conceptual Framework” in COCA Academic

1	2019	ACAD	European Research Studies				ISBN 978-87- 7681-491-5 boobkboon.com. # White, L.R. & Clinton, B.D. 2014. The Conceptual Framework for Managerial Costing. Institute of
2	2019	ACAD	Survey Methodology				future research. # 2 Quantile regression imputation for complex survey data # Consider a conceptual framework in which samples are drawn
3	2019	ACAD	Archives of Public Health				income countries remains a challenge. The United Nations Children's Fund (UNICEF) conceptual framework provides the best opportunity to
4	2019	ACAD	PLoS ONE				structure and fish assemblages in the Mulgrave River, north-east Queensland: towards a new conceptual framework for understanding fish-h
5	2019	ACAD	Journal of Shipping and Trade				Grewal, Devinder: Selecting the location of distribution centre in logistics operations: A conceptual framework and case study. vol. 17, issue 3
6	2019	ACAD	Asia-Pacific Science Education				). doi: **28;260;TOOLONG # Kelley, TR, Knowles, JG: A conceptual framework for integrated STEM education. vol. 3, issue 1, pp. 11.
7	2019	ACAD	City, Territory and Architecture				6: 59 # Bibri SE (2019e) Data-driven smart sustainable cities: a conceptual framework for urban intelligence functions and related processes,
8	2019	ACAD	Harvard J Law Public Policy				Limits # All governments negotiate the balance between permitting and restricting speech within an overarching conceptual framework of th
9	2019	ACAD	Harvard J Law Public Policy				decisions do showcase the Court's enduring embrace of second-wave free speech expansionism- the dominant conceptual framework for ro
10	2019	ACAD	Harvard J Law Public Policy				. This Article, however, focuses on the nature of the changes to the conceptual framework of free speech as manifested in legal, and primarily

8.2.2 Implications for EAP material designers

One major implication for EAP material designers is the need to incorporate more academic collocations into textbooks. This is an important condition for incidental and direct learning to happen. There is a consensus in the literature that collocations can be acquired incidentally through reading and listening, given sufficient exposure (Ellis, 1997; Nation, 2013; Siyanova-Chanturia & Schmitt, 2008; Webb & Chang, 2020). As pointed out in the discussion in Chapter 7, collocations do not seem to receive much attention compared to other aspects of vocabulary in EAP textbooks (Coxhead et al., 2020). Results from the evaluation of the ACL (Ackermann & Chen, 2013) and the AECL (Lei & Liu, 2018) (see Chapter 4) suggest some useful sources for material developers to consider and we turn to those points now.

One of the findings of this study suggests that the most frequent 500 academic lexical collocations of either the ACL or the AECL could be an immediate and practical target of EAP programs. These collocations are included in Appendix H. It could be argued that learners can naturally be exposed to high frequency vocabulary, meaning that these items might not require explicit instruction. That said, the frequency of academic collocations tends to be much lower than that of single academic words. Even when learners know two words individually, they may not know that they collocate (as discussed in Chapter 7). Therefore, these collocations still deserve attention from EAP educators. Material designers could make use of these sources to design textbooks or materials for EAP programs.

The overlapping list of 1,298 academic collocation items appearing in both the ACL and the AECL is another useful source of academic collocations. The full list can be found in Appendix I. Although the development of the ACL and the AECL was based on different corpora, criteria and procedures (see Chapter 4), these items occur in both lists, indicating that they are likely important and useful for EAP. Textbook designers could consider using this overlapping list to develop materials for EAP learners.

8.2.3 Implications for test developers

This study also has several implications for test developers. First, recognition and recall knowledge of academic collocations are two separate constructs, which is why different measures were needed for this study to assess each type of knowledge. The findings of the present study highlighted this difference and revealed that learners had better recognition knowledge of academic collocations than their recall knowledge (see Chapter 5). If only one test was used to assess learners' knowledge, the results would have been misleading. The implication here is that it is important for test developers to clarify the construct being measured (either recognition or recall) to provide a more nuanced indication of learners' vocabulary knowledge.

Second, it is critical for test developers to clearly understand how a word list was made and evaluated if they plan to use it for testing. This is because the way a list is created has an impact on the test. If the list is not well made, the test items may not properly represent the target knowledge being measured. Depending on the wordlist evaluation results, a list can be used as how it is or it can be

adapted to match the assessment purpose. Taking the ACL (Ackermann & Chen, 2013) used in this study as an example, the list was narrowed down first to remove items that were more general than academic (e.g., *make contact*). The remaining items were then used for the development of the ACTs. The final aim was to select the representative items in the list.

Third, this study demonstrated the advantages of using test development and validation frameworks. The evidence-centred design (Mislevy & Yin, 2013) allows test developers to carefully consider every aspect of test development to strengthen the test design. Kane's (2004, 2013) argument-based approach then helps to add rigorous evidence to the quality of tests. The key point about the argument-based approach is being able to make justifiable claims about the test scores and specifying which evidence is needed for those claims. For example, this study avoided making claims such as the efficacy of the ACTs over other measures of proficiency to provide an indication of learners' general language proficiency. Instead, the present study only claimed that scores on the ACTs were related to scores on other tests measuring similar constructs, such as IELTS. The evidence found to support this claim was the moderate correlation between scores on the ACTs and IELTS scores.

Finally, the evidence from this study suggests that the post-test interview is a useful tool to provide insight into the testing process. Looking closely into the thinking process of test-takers helps test developers to discover whether a test has acted as intended. Especially, this study suggests that investigating test-taking strategies is a useful method to shed light on the validity of test results. The strategies revealed by test-takers help to determine whether the produced answers are based on their actual knowledge or other external factors. For example, if the strategy of guessing is employed for the AC Recognition Test or dictionary use for the AC Recall Test, the validity of the test results will be threatened because there is a mismatch between what the tests are designed to elicit and what test-takers produce. The post-test interview is especially beneficial for online assessment in which the test-taking process cannot be directly observed. The interview questions in this study (see Appendix B) are examples of how different aspects of the tests and the test-taking process could be elicited. Analysing participants' opinions and reflections helps test developers gain more evidence for the assessment of the test quality.

8.2.4 Implications for word list developers

For word list developers, an important implication is that careful consideration of the definition of academic multiword units, especially collocations, is needed. Units of counting of single words, such as types and lemmas, can be difficult to consistently apply to long lists of collocations. For some kinds of collocations, such as verb + noun or adjective + noun, collocational patterns do not change when used with different verb forms (e.g., *make/ makes/ made /making decision*) or with singular or plural nouns (e.g., *living condition/ conditions*). This suggests that lemmas are appropriate as a unit of counting for word lists in this instance. For other kinds of collocations, such as adverb + verb, types might be the better unit of counting. For example, *widely* frequently occurs with *known/ distributed/ accepted* but not with other types of verbs such as *know/ distribute/ accept*. Therefore, clearly establishing the unit of counting at the early stage of developing a list is necessary.

In addition, the present study has demonstrated that word list evaluation is an important step in word list development. List users need to know what decisions have been made during the development of word lists and why. It is important that wordlist research is well documented and transparent regarding the process of developing and validating lists so that users can select the most suitable list for their needs.

8.3 Research limitations

This study has a number of limitations. One drawback of this research is related to test administration. There were some interruptions during the test-taking process via the online platform, and some participants consulted dictionaries which was out of my control. It was not possible to individually observe test takers over Zoom under lockdown. In addition, when the two academic collocation tests were delivered in a sequence, a priming effect could not be completely ruled out. These elements might have affected the test results in the present study to a certain extent, as thoroughly discussed in Chapter 7 (Section 7.3.3).

In addition to the limitations caused by external factors, constraints of the ACTs should be acknowledged. First, the test formats did not accurately reflect real tasks in the academic context. As discussed in Chapter 2 (Section 2.4), in the real-world context in which academic collocations are employed, learners are unlikely to be given the options to choose from or the initial letters to recall a

certain collocation. Similarly, the simplified context sentences in the ACTs did not exactly reflect the academic language domain for which the ACTs were intended, despite efforts to keep them as formal and academically relevant as possible. Second, limitations of the ACL (Ackermann & Chen, 2013) applied to the ACTs, as discussed in Chapter 7 (Section 7.2.2). For example, the ACTs do not cover grammatical collocations (e.g., *lack of* or *based on*) that can also be useful for pedagogy. That said, compared to previous measures of collocation knowledge, the ACTs reflect a wide range of collocation kinds. While most of the prior tests consist of one or two collocation kinds (e.g., Gyllstad, 2009; Nguyen & Webb, 2017; Voss, 2012), the ACTs contain eight different kinds of collocations that appear in the ACL (Ackermann & Chen, 2013). Lastly, the AC Recognition Test in its current format was unable to capture the knowledge of learners with a wide range of proficiency. In other words, this test lacks items of higher difficulty to measure advanced learners' knowledge of academic collocations.

There are also limitations that need to be noted about the wordlist evaluation. Reporting on frequency information of academic collocations in this study did not take different forms of the same collocation into account (see Chapter 3, Section 3.1.1). This constraint was related to the analysis tool. Antconc software (Anthony, 2019) was used to process the corpora and provide frequency information of the items in their fixed forms in the source list. The frequency of different forms could only be searched and calculated manually item by item in both academic and fiction corpora. This means that the task could not be completed given the huge number of items in the two academic collocation lists.

8.4 Future research directions

The findings and limitations of the present study suggest directions for future research for three different areas: testing, validation and wordlist evaluation.

8.4.1 Directions for research on testing knowledge of academic collocations

One avenue for future testing research would be conducting an experimental study in which the ACTs are used in class along with the ACL (Ackermann & Chen, 2013). The focus of such research would be to determine the effectiveness of an intervention (e.g., a teaching method or material enhancement) on learners' academic collocation knowledge. This is different from the "cold testing"

context in this thesis which did not assume that test-takers were familiar with the term “academic collocations”; therefore, “academic phrases” or “phrases in academic context” were used in the test instruction for clarification. With experimental studies, learners would have time to learn explicitly or implicitly the academic collocations such as those in the list of Ackermann and Chen (2013) before being tested. The ACTs could be delivered before and after an intervention to estimate the learning gain of academic collocations. This kind of research could strengthen the connection between teaching and testing, which is a major advantage of wordlist-based tests.

Another fruitful area for further work would be to investigate potential factors that may affect knowledge of academic collocations, such as L1 congruency, semantic opacity or strength of association. While the current research found that the frequency of academic collocations as whole units had a negligible correlation with learners’ knowledge, it would be interesting to find out if the frequency of collocation components (i.e., individual words in the collocation pair) has a greater impact on the learning acquisition. To this end, Appendix J contains the frequency information needed to carry out this future research and for interest. Furthermore, identifying which of those is a determining factor for the development of academic collocation knowledge would provide useful implications for pedagogy.

8.4.2 Directions for validation research of academic collocation tests

A natural progression of this study is to add the Utilisation inference into the validation framework of the ACTs and examine whether the test scores are valid for application in a specific context. For instance, the test scores could be used to make a decision on whether learners would benefit from a supplementary course on academic writing to increase their awareness, practice and use of academic collocations. This inference would then require evidence that a higher test score on the ACTs corresponds to better writing skills. A cut-off score needs to be identified for determining whether learners need extra support. In order to achieve a robust cut-off score, the ACTs should be administered to a large number of learners and the test scores could be analysed in connection with the quality of learners’ writing pieces. Given the fact that the ACTs have not been employed in any course or program, further research is needed to explore the uses of the test scores.

A further study could also replicate and expand on the current research. As the test administration in the present study reveals some weaknesses (Section 8.3), future research could employ face-to-face testing instead of an online method and compare the test results to those in this thesis. Such a replication study would provide a good opportunity to revisit any misfitting items detected in this study and determine whether and how they are problematic. It has been suggested in the current study that those misfits could be results of a random effect in the statistical model. Another possible extension would be to make use of other useful collocations in the lists of Ackermann and Chen (2013) to create multiple versions of the ACTs. This would allow for the assessment of a wider range of academic collocations. The test development and validation have been well documented in this thesis. Hence, these directions for replication studies would be practical and beneficial. Finally, future research could also re-examine the difficulty level of the AC Recognition Test by administering the test to participants at a wider range of proficiency levels, and trialling a different test format without repeating one common word among the provided options. Such a study would provide more evidence for the validity assessment of the AC Recognition Test in the present study.

8.4.3 Directions for research on wordlist evaluation

Future research could focus on evaluating other lists of multiword units such as Simpson-Vlach and Ellis (2010) and Rogers et al. (2021). These lists are created for academic purposes, but they have not been evaluated by an independent study as the current research did with the two lists of academic collocations. The evaluation could start with a particular purpose in a specific context. Apart from the aim of developing tests as in the present study, other purposes such as teaching and learning vocabulary or designing materials would also benefit from such wordlist evaluation studies.

Further research might also focus on developing a more accessible evaluation framework for classroom teachers. The coverage comparison and Nation's (2016) evaluation framework used in this study require specialised resources and knowledge, making them perhaps more suitable for researchers rather than teachers. A more simplified framework would be beneficial for teachers to determine the better list for their needs, especially when research on evaluating word lists is still rare.

8.5 Reflections on my PhD journey

Assessing collocation knowledge is a research area that has piqued my interest since my MA study, and I wanted to further explore this area during my PhD journey. Although I started my PhD with some background in collocation research, I have gained new knowledge when conducting this PhD project. First, I challenged myself by developing two tests on recognition and recall knowledge of academic collocations instead of just the recognition test as in the initial plan, and this was a rewarding experience. The recall test was much more difficult to design than the recognition one, especially because the testing context involved learners with different L1 backgrounds. The creation of the AC Recall Test is a testament to my effort to overcome the obstacles and my enthusiasm to contribute to the research literature. Additionally, through the comparison between recognition and recall knowledge of academic collocations, I discovered that academic collocations, although straightforward in recognition, are challenging in recall. This explains why, as a proficient English user, I have no difficulty in reading academic texts, but occasionally produce deviant academic collocations. Developing the ACTs provided me with the chance to acquire new collocations and apply them in my own academic writing. This valuable experience will certainly support my teaching career in helping learners to develop their knowledge of academic collocations.

Second, I have improved my expertise in language testing, especially in validation research. The argument-based validation approach (Kane, 2004, 2013) has enabled me to better understand the logic of a validation process and apply it to my own research. The process of collecting and analysing research data fostered my ability to work with both quantitative and qualitative data. This mixed-methods approach offered me an opportunity to learn new skills that will facilitate my future research practice and help me to become an independent researcher.

Last but not least, wordlist evaluation was a completely new area into which I delved for the first time. Through the evaluation process of two academic collocation lists, I recognised corpus-based word lists are valuable resources and I sought to connect these research outcomes to vocabulary instruction and assessment. Within the scope of this study, I have only focused on the assessment direction through the development of the ACTs using the ACL (Ackermann & Chen, 2013). I hope to further investigate the other direction in using word lists in the classroom context in the future as a teacher researcher.

Overall, my PhD journey has been challenging but rewarding. Through this work, I have learned new research methods and skills, had the opportunity to contribute to the research literature and developed into both a teacher with a stronger research background as well as a researcher exploring more areas in the field of applied linguistics.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Ackermann, K., De Jong, J. H. A. L., Kilgarriff, A., & Tugwell, D. (2011). The Pearson International Corpus of Academic English (PICAE). *Corpus Linguistics*.

Adolphs, S., & Durow, V. (2004). Social-cultural integration and the development of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 107–126). John Benjamins Publishing Company.

Alavi, S. M., Kaivanpanah, S., & Masjedlou, A. P. (2018). Validity of the listening module of international English language testing system: Multiple sources of evidence. *Language Testing in Asia*, *8*(1), 1–17. https://doi.org/10.1186/s40468-018-0057-4

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173–195. https://doi.org/10.1093/applin/22.2.173

Anthony, L. (2019). *AntConc* (3.5.8) [Computer software]. Waseda University. http://www.laurenceanthony.net/software

Aryadoust, V. (2011). Validity arguments of the speaking and listening modules of international English language testing system: A synthesis of existing research. *Asian ESP Journal*, *7*(2), 28–54.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1–34.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System*, *21*(1), 101–114. https://doi.org/10.1016/0346-251X(93)90010-E

Ballance, O. J., & Coxhead, A. (2020). How much vocabulary is needed to use a concordance? *International Journal of Corpus Linguistics*, *25*(1), 36–61. https://doi.org/10.1075/ijcl.17116.bal

Balotsky, E. R., Stagliano, A. J., & Haub, E. K. (2016). How accreditation engenders pedagogical improvements through the assurance of learning process-a case study from a strategic management course. *Academy of Business Journal, 1*(1), 75–95.

Barfield, A. (2009). Exploring productive L2 collocation knowledge. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners: Papers and perspectives in honour of Paul Meara* (pp. 95–110). Multilingual Matters.

Barfield, A., & Gyllstad, H. (2009). *Researching collocations in another language: Multiple interpretations*. Palgrave Macmillan.

Begagić, M. (2015). English language students' productive and receptive knowledge of collocations. *Explorations in English Language and Linguistics*, *2*(1), 46–67.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins Publishing Company.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). Longman.

BNC Consortium. (2007). *The British National Corpus, XML Edition*. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2554

Boers, F. (2021). *Evaluating second language vocabulary and grammar instruction: A synthesis of the research on teaching words, phrases, and patterns*. Routledge.

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb–noun collocations. *Language Teaching Research*, *18*(1), 54–74.

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

Bonk, W. J. (2001). Testing ESL learners' knowledge of collocations. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (pp. 113–142). University of Hawaii Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 2. Research designs: Quantitative, qualitative, neuropsychological, and biological.* (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

British Council. (n.d.). *IELTS for Vietnam*. IELTS Asia. Retrieved September 13, 2021, from https://www.ieltsasia.org/vn/en/study-in-vietnam

Brown, D. (2018). *Developing a measure of L2 learners' productive knowledge of English collocations* [Unpublished doctoral dissertation]. Cardiff University.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Bui, T., Boers, F., & Coxhead, A. (2020). Extracting multiword expressions from texts with the aid of online resources: A classroom experiment. *ITL - International Journal of Applied Linguistics, 171*(2), 221–252. https://doi.org/10.1075/itl.18033.bui

Cao, D., & Badger, R. (2021). Cross-linguistic influence on the use of L2 collocations: the case of Vietnamese learners. *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2020-0035

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing*, *29*(1), 19–27. https://doi.org/10.1177/0265532211417211

Chapelle, C. A. (2020). Validity in language assessment. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 11–20). Routledge.

Chapelle, C. A., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Chapelle, C. A., & Lee, H. (2021). Understanding argument-based validity in language testing. In C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (pp. 19–44). Cambridge University Press. https://doi.org/10.1017/9781108669849.004

Chen, I. (2019). *A corpus-driven receptive test of collocational knowledge* [Unpublished doctoral dissertation]. University of Melbourne.

Chon, Y., & Shin, D. (2013). A corpus-driven analysis of spoken and written academic collocations. *Multimedia-Assisted Language Learning*, *16*(3), 11–38.

Chung, Y.-R. (2014). *A test of productive English grammatical ability in academic writing: Development and validation* [Unpublished doctoral dissertation]. Iowa State University.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cohen, A. D. (1998). *Strategies in learning and using a second language*. Longman.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, *3*(4), 307–331.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

Coxhead, A. (2008). Phraseology and English for academic purposes. Challenges and opportunities. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 149–161). John Benjamins Publishing Company.

Coxhead, A. (2018). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives*. Routledge.

Coxhead, A. (2019). Analysis of corpora. In *The Routledge handbook of research methods in applied linguistics* (pp. 464–473). Routledge. https://doi.org/10.4324/9780367824471-39

Coxhead, A., & Dang, T. N. Y. (2019). Vocabulary in university tutorials and laboratories: Corpora and word lists. In K. Hyland & L. L. C. Wong (Eds.), *Specialised English: New directions in ESP and EAP research and practice*. Routledge.

Coxhead, A., Dang, T. N. Y., & Mukai, S. (2017). Single and multi-word unit vocabulary in university tutorials and laboratories: Evidence from corpora and textbooks. *Journal of English for Academic Purposes*, *30*, 66–78. https://doi.org/10.1016/j.jeap.2017.11.001

Coxhead, A., Rahmat, Y., & Yang, L. (2020). Academic single and multiword vocabulary in EFL textbooks: Case studies from Indonesia and China. *TESOLANZ Journal*, *28*, 75–88.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*(5), 570–590. https://doi.org/10.1093/applin/amt056

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, *67*(4), 959–997. https://doi.org/10.1111/lang.12253

Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL - International Journal of Applied Linguistics*, *167*(2), 132–158. https://doi.org/10.1075/itl.167.2.02dan

Davies, M. (2008-2017). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-2017*. Available online at https://www.english-corpora.org/coca/.

Dokchandra, D. (2019). Exploring Thai university students' receptive and productive knowledge of collocations across four academic faculties. *Advances in Language and Literary Studies*, *10*(6), 115–123.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157–169.

Dursun, A., & Li, Z. (2021). A systematic review of argument-based validation studies in the field of language testing (2000–2018). In C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (pp. 45–70). Cambridge University Press. https://doi.org/10.1017/9781108669849.005

Ellis, R. (1997). *SLA research and language teaching*. Oxford University Press.

Erten, I. H., & Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semantically unrelated sets. *System*, *36*, 407–422.

University of Cambridge (2005). *Cambridge IELTS 4: Examination papers from University of Cambridge ESOL Examinations*. Cambridge University Press.

Eyckmans, J. (2009). Toward an assessment of learners' receptive and productive syntagmatic knowledge. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 139–152). Palgrave Macmillan UK. https://doi.org/10.1057/9780230245327_11

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques* (pp. 83–102). Routledge.

Fernández, B. G., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? *ITL-International Journal of Applied Linguistics*, *166*(1), 94–126.

Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*.

Fisher, W. (1992). Reliability statistics. *Rasch Measurement Transactions*, *6*(3), 238.

Frankenberg-Garcia, A. (2018). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes*, *35*, 93–104. https://doi.org/10.1016/j.jeap.2018.07.003

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Gaballa, H. E.-B. M., & Al-Khayri, M. A. (2014). Testing collocational knowledge of Taif University English seniors. *IOSR Journal of Humanities and Social Science*, *19*(11), 63–90.

Gablasova, D., Brezina, V., & Mcenery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(S1), 155–179. https://doi.org/10.1111/lang.12225

Gardner, D. (2007). Validating the construct of "word" in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241–265. https://doi.org/10.1093/applin/amm010

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327. https://doi.org/10.1093/applin/amt015

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, *19*(6), 645–666. https://doi.org/10.1177/1362168814559798

Garnier, M., & Schmitt, N. (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System*, *59*, 29–44. https://doi.org/10.1016/j.system.2016.04.004

Gitsaki, C. (1999). *Second language lexical acquisition: A study of the development of collocational knowledge*. International Scholars Publications.

Granger, S., & Larsson, T. (2021). Is core vocabulary a friend or foe of academic writing? Single-word vs multi-word uses of THING. *Journal of English for Academic Purposes*, 100999. https://doi.org/10.1016/j.jeap.2021.100999

Gyllstad, H. (2009). Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 153–170). Palgrave Macmillan. https://doi.org/10.1057/9780230245327_12

Gyllstad, H., & Schmitt, N. (2018). Testing formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 174–191). Routledge.

Hartshorn, J., & Hart, J. (2016). Comparing the Academic Word List with the Academic Vocabulary List: Analyses of frequency and performance of English language learners. *The Journal of Language Teaching and Learning*, *6*(2), 70–87.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development–a progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 29–56). John Benjamins Publishing Company.

Henriksen, B., & Westbrook, P. (2017). Responding to research challenges related to studying L2 collocational use in professional academic discourse. *Vocabulary Learning and Instruction*, *6*(1), 32–47.

Hill, J., & Lewis, M. (1997). *LTP dictionary of selected collocations*. Language Teaching Publications.

Hines, S. (2010). Evidence-centered design: The TOEIC® speaking and writing tests. *TOEIC Compendium*, 1–31.

Hinkel, E. (Ed.). (2018). *Teaching essential units of language: Beyond single-word vocabulary*. Routledge.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.

Howarth, P. A. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Max Niemeyer.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41–62. https://doi.org/10.1111/j.1473-4192.2008.00178.x

Hyland, K., & Shaw, P. (2016). Introduction. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 1–13). Routledge. https://doi.org/10.4324/9781315657455

Ishikawa, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N., Tono, Y., & Murata, M. (2003). JACET8000: JACET list of 8000 basic words. *Tokyo: JACET*, *3*.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, *2*(3), 135–170.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). American Council on Education and Praeger.

Kane, M. (2011). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17. https://doi.org/10.1177/0265532211417210

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17.

Keshavarz, M. H.-S., & Salimi, H. (2007). Collocational competence and cloze test performance: A study of Iranian EFL learners. *International Journal of Applied Linguistics*, *17*(1), 81–92. https://doi.org/10.1111/j.1473-4192.2007.00134.x

Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography*, *24*(1), 29–49. https://doi.org/10.1093/ijl/ecq039

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1191/026553299672614616

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lee, S., & Shin, S.-Y. (2021). Towards improved assessment of L2 collocation knowledge. *Language Assessment Quarterly*, 1–26. https://doi.org/10.1080/15434303.2021.1908295

Lei, L., & Liu, D. (2018). The academic English collocation list. *International Journal of Corpus Linguistics*, *23*(2), 216–243. https://doi.org/10.1075/ijcl.16135.lei

Li, J., & Schmitt, N. (2010). The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach. In *Perspectives on formulaic language: Acquisition and communication* (pp. 22–46). Continuum.

Linacre, J. M. (1997). KR-20/Cronbach alpha or Rasch person reliability: Which tells the "truth." *Rasch Measurement Transactions*, *11*(3), 580–581.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2019). *Winsteps® Rasch measurement computer program user's guide*. www.winsteps.com/winman/webpage.htm

Liu, D., & Zhong, S. (2016). L2 vs. L1 use of synonymy: An empirical study of synonym use/acquisition. *Applied Linguistics*, *37*(2), 239–261. https://doi.org/10.1093/applin/amu022

Macis, M., & Schmitt, N. (2017). Not just 'small potatoes': Knowledge of the idiomatic meanings of collocations. *Language Teaching Research*, *21*(3), 321–340. https://doi.org/10.1177/1362168816645957

Martinez, R. (2011). *The development of a corpus-informed list of formulaic sequences for language pedagogy* [Unpublished doctoral dissertation]. University of Nottingham.

Martinez, R. (2019). Integrating corpus tools into mixed methods research. In *The Routledge handbook of research methods in applied linguistics* (pp. 211–229). Routledge. https://doi.org/10.4324/9780367824471-19

Miller, D., & Biber, D. (2015). Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition. *International Journal of Corpus Linguistics*, *20*(1), 30–53.

Miller, J. (2020). The bottom line: Are idioms used in English academic speech and writing? *Journal of English for Academic Purposes*, *43*. https://doi.org/10.1016/j.jeap.2019.100810

Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36*(8), 463–469. https://doi.org/10.3102/0013189X07311660

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to evidence-centered design. *ETS Research Report Series*, *1*, i–29.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Mislevy, R. J., & Yin, C. (2013). Evidence-centered design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 208–222). Routledge.

Nation, I. S. P. (2000). Learning vocabulary in lexical sets: Dangers and guidelines. *TESOL Journal*, *9*(2), 6–10. https://doi.org/10.1002/j.1949-3533.2000.tb00239.x

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–14). John Benjamins Publishing Company.

Nation, I. S. P. (2012). *The BNC/COCA word family lists*. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nation, I. S. P., Shin, D., & Grant, L. (2016). Multiword units. In *Making and using word lists for language learning and testing* (pp. 71–79). John Benjamins Publishing Company.

Nation, I. S. P., & Sorell, J. (2016). Corpus selection and design. In *Making and using word lists for language learning and testing* (pp. 95–105). John Benjamins Publishing Company.

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge University Press.

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge University Press.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins Publishing Company.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing*, *6*(2), 199–215.

Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3), 298–320.

Oh, S. R. (2018). *Investigating test-takers' use of linguistic tools in second language academic writing assessment* [Unpublished doctoral dissertation]. Teachers College, Columbia University.

Oliveros, J. C. (2015). *Venny. An interactive tool for comparing lists with Venn's diagrams* (2.1.0) [Computer software]. http://bioinfogp.cnb.csic.es/tools/venny/index.html

Oxford University Press. (n.d.). *Oxford Online Placement Test*. https://www.oxfordenglishtesting.com

Pearlman, M. (2008). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language*^{TM} (pp. 227–258). Routledge.

Pecorari, D., Shaw, P., & Malmström, H. (2019). Developing a new academic vocabulary test. *Journal of English for Academic Purposes*. https://doi.org/10.1016/j.jeap.2019.02.004

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institute.

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Rogers, J., Müller, A., Daulton, F. E., Dickinson, P., Florescu, C., Reid, G., & Stoeckel, T. (2021). The creation and application of a large-scale corpus-based academic multi-word unit list. *English for Specific Purposes*, *62*, 142–157. https://doi.org/10.1016/j.esp.2021.01.001

Salehi, M. (2011). Test taking strategies: Implications for test validation. *Journal of Language Teaching and Research*, *2*(4), 850–858.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Schmitt, N. (2013). Formulaic language and collocation. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). Wiley-Blackwell.

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Shin, D., & Nation, I. S. P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, *62*(4), 339–348. https://doi.org/10.1093/elt/ccm091

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512. https://doi.org/10.1093/applin/amp058

Siyanova-Chanturia, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*. https://doi.org/10.3138/cmlr.64.3.429

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, *36*(5), 549–569.

Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (Eds.). (2019). *Understanding formulaic language: A second language acquisition perspective*. Routledge.

Smith Jr, E. V. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, *6*(2), 147.

Sonbul, S., Salam El-Dakhs, D. A., & Al-Otaibi, H. (2020). Productive versus receptive L2 knowledge of polysemous phrasal verbs: A comparison of determining factors. *System, 95*. https://doi.org/10.1016/j.system.2020.102361

Sorell, C. J. (2013). *A study of issues and techniques for creating core vocabulary lists for English as an international language* [Unpublished doctoral dissertation]. Victoria University of Wellington.

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition, 43*(1), 181–203.

Stoeckel, T.-B. (2016). Is "I Don't Know" a viable answer choice on the Vocabulary Size Test? *TESOL Quarterly*, *50*(4), 965–975. https://doi.org/10.1002/tesq.325

Thompson, P., & Nesi, H. (2001). The British Academic Spoken English (BASE) corpus project. *Language Teaching Research*, *5*(3), 263–264.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press.

Tran, D. (2020). *Argument-based validation of a high-stakes Listening test in Vietnam* [Unpublished doctoral dissertation]. Victoria University of Wellington.

Trang, N. Q. (2010). Difficulties experienced by Vietnamese lecturers teaching IELTS speaking at university level and some suggested solutions. *VNU Journal of Foreign Studies, 26*(4).

Vaseghi, R., Rad, Z. B., & Azarfam, A. Y. (2020). Connecting formulaic sequences and moves in applied linguistics research article results. *Journal of English Language Pedagogy and Practice, 13*(27), 53–71.

Vilkaitė-Lozdienė, L. (2016). Formulaic language is not all the same: Comparing the frequency of idiomatic phrases, collocations, lexical bundles, and phrasal verbs. *Taikomoji Kalbotyra*, *8*. https://doi.org/10.15388/TK.2016.17505

Vilkaitė-Lozdiene, L., & Schmitt, N. (2019). Frequency as a guide for vocabulary usefulness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 81–96). Routledge.

Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design* [Unpublished doctoral dissertation]. Iowa State University.

Vu, D., & Michel, M. (2021). An exploratory study on the aspects of vocabulary knowledge addressed in EAP textbooks. *Dutch Journal of Applied Linguistics*.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52. https://doi.org/10.1017/S0272263105050023

Webb, S. A., & Chang, A. C.-S. (2012). Second language vocabulary growth. *RELC Journal*, *43*(1), 113–126. https://doi.org/10.1177/0033688212439367

Webb, S. A., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., & Chang, A. C.-S. (2020). How does mode of input affect the incidental learning of collocations? *Studies in Second Language Acquisition*, 1–22.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

West, M. (1953). *A general service list of English words*. Longman.

Wongkhan, P., & Thienthong, A. (2020). EFL learners' acquisition of academic collocation and synonymy: Does their academic experience matter? *RELC Journal*. https://doi.org/10.1177/0033688219895046

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

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

Wright, B. D., & Tennant, A. (1996). Sample size again. *Rasch Measurement Transactions*, *9*(4), 468.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer Singapore.

Wulff, S. (2019). Acquisition of formulaic language from a usage-based perspective. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 19–37). Routledge. https://doi.org/10.4324/9781315206615-2

Youn, S. J. (2013). *Validating task-based assessment of L2 pragmatics in interaction using mixed methods* [Unpublished doctoral dissertation]. University of Hawai'i at Manoa.

Zhang, X. (2013). The "I don't know" option in the Vocabulary Size Test. *TESOL Quarterly*, *47*(4), 790–811. https://doi.org/10.1002/tesq.98

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

Appendices

Appendix A

Background Questionnaire

BACKGROUND QUESTIONNAIRE

Welcome to the Examining English learners' vocabulary knowledge research project!

First of all, thank you for considering taking part in this research project. Before you begin, please answer the two screening questions below to make sure that you belong to the target group that we are looking for. If you are eligible for this study and wish to take part, please continue. If you are not eligible, thank you for your time. The survey will be terminated.

Are you a non-native English speaker student at Victoria University of Wellington?

◯ Yes ◯ No

Have you been studying in an English-speaking country for at least 12 months?

◯ Yes ◯ No

In this section, I would like you to answer a few questions about your personal and language learning background. This information will help. Please provide the following information by selecting the appropriate response or typing your response in the space.

What is your name?

_____

What is your gender?

◯ Male ◯ Female ◯ Other

How old are you? (Please fill in a number)

_____

What is your first language?

_____

Have you got any English language certificate such as IELTS or TOEFL?

◯ Yes ◯ No

Please provide the name and overall score in the English language certificate that you have. For example, IELTS 6.5

_____

What study level are you in at the moment?

◯ First year undergraduate

◯ Second year undergraduate

◯ Third year undergraduate

◯ Fourth year undergraduate

◯ Other (Please specify) _____

What is your major?

_____

How long have you been learning English?

_____ year(s) _____ month(s)

How many years/ months in total have you been studying in an English-speaking country?

_____ year(s) _____ month(s)

Appendix B

Post-Test Interview

A. Introduction

1. Greet the interviewee. Explain the aim and the estimated time of the interview.
2. Remind the interviewee of the three tests s/he already took.

B. Reflection

1. In your opinion, which test was the easiest? Which test was the most difficult?
2. Do you have any comments on the tests?
3. Did you have any difficulty in understanding the context sentences and the meaning of the phrases in the brackets?
4. Did you use dictionary or any external sources when taking the tests?
5. When you left any item blank, does that mean you did not know the answer, or you just skipped it?
6. Could you remember anything in the gap-filling test when taking the last multiple-choice test?
7. Did your memory help you answer any items in the last multiple-choice test?
8. Did you do all the tests at once or stop and then come back?

C. Test-retaking

Questions for further explanation:

- Why did you choose/ have this answer?
- Did you consider any other options?
- Were you confident with this answer?
- Had you met this phrase before?

Appendix C

The Academic Collocation Recognition Test

This is the Recognition Test of academic collocations. It tests your ability to recognize common written phrases in academic contexts such as in university textbooks or in research papers. There are 60 items in this test.

Please choose the best academic phrases (A, B, C or D) to fill in each of the following blanks.

Example:

0. The region in a city is called a(n) _____ _____.
   A. certain area
   B. remote area
   (C) urban area
   D. rural area

1. The two approaches have a _____ _____. One uses interviews while the other does not.
   A. significant difference
   B. significant effect
   C. significant finding
   D. significant number
2. Parents also should encourage their children to get involved in a _____ _____ such as swimming or running.
   A. creative activity
   B. mental activity
   C. physical activity
   D. social activity
3. One of the tips to stay healthy is eating a _____ _____ of fruits and vegetables.
   A. wide area
   B. wide band
   C. wide range
   D. wide scope
4. An organization was set up to _____ the _____ of traffic jams.
   A. address the barrier
   B. address the gap
   C. address the issue
   D. address the limitation
5. People may not want to share their personal information including age, salary and _____ _____.
   A. religious belief
   B. religious freedom
   C. religious liberty
   D. religious right

6. A set of ideas or plans that a country uses as a basis for making decisions is called _____ _____.
   A. government agency
   B. government official
   C. government policy
   D. government reform
7. An unwritten rule of manners that are considered acceptable in a group or society is called a _____ _____.
   A. social life
   B. social network
   C. social status
   D. social norm
8. Customer spending was the most important _____ _____ behind the economic growth in the summer.
   A. creative force
   B. driving force
   C. excessive force
   D. physical force
9. The two conditions cannot happen at the same time. They are _____ _____.
   A. mutually acceptable
   B. mutually beneficial
   C. mutually exclusive
   D. mutually supportive
10. Everything in that event was _____ _____, including the location, the guest list and the speeches.
    A. well documented
    B. well established
    C. well illustrated
    D. well equipped
11. The interviews were given to a(n) _____ _____ of students. Any student was able to join if they wanted.
    A. current sample
    B. modest sample
    C. overall sample
    D. random sample
12. The _____ _____ of this program is to prevent further damage to the sea.
    A. ultimate cost
    B. ultimate form
    C. ultimate goal
    D. ultimate price
13. The math teacher _____ a(n) _____ for his students who missed the final exam to take the test again.
    A. offered an advantage

B. offered a benefit
C. offered an opportunity
D. offered a reward

14. Research has found that students' grades and their relationship with parents are _____ _____.
A. positively affected
B. positively influenced
C. positively predicted
D. positively associated

15. The event welcomed a(n) _____ _____ of guests from over 30 countries.
A. discrete group
B. ethnic group
C. racial group
D. diverse group

16. Research results show that (a) _____ _____ of smokers tend to be drinkers also.
A. high performance
B. high ability
C. high incidence
D. high percentage

17. The university is highly popular and runs various programs in a wide number of _____ _____.
A. academic achievements
B. academic disciplines
C. academic performances
D. academic standards

18. Science is an _____ _____ of society.
A. essential component
B. essential function
C. essential quality
D. essential condition

19. There is more work that needs to be done to _____ _____ the effect of this change.
A. fully satisfy
B. fully expect
C. fully support
D. fully understand

20. Many _____ _____ can affect your business such as the economic, political and social environment of the locations where the company operates.
A. interpersonal factors
B. external factors
C. internal factors
D. subjective factors

21. The library and the union building are located in _____ _____.
A. close attention

B. close contact
C. close encounter
D. close proximity

22. The difference in wages _____ _____, neither increasing nor reducing over time.
A. remained active
B. remained constant
C. remained neutral
D. remained viable

23. An increase in the use of private vehicles _____ a _____ to the environment.
A. confronts a threat
B. poses a threat
C. reduces a threat
D. counters a threat

24. It will be harder to get the job without _____ _____ in the field.
A. daily experience
B. human experience
C. prior experience
D. whole experience

25. The _____ _____ of a dollar bill is only that of a piece of paper.
A. aesthetic value
B. intrinsic value
C. monetary value
D. sentimental value

26. The government tried to prevent the _____ _____ of the new drug because of its side effects.
A. efficient use
B. practical use
C. widespread use
D. proper use

27. Water pollution leads to a _____ _____ of disease.
A. high incidence
B. high priority
C. high quality
D. high standard

28. Short sentences should be used when _____ a _____, such as the instructions on how to build a model.
A. repeating a process
B. describing a process
C. permitting a process
D. reversing a process

29. Air pollution has become the _____ _____ that affects the development of tourism in this city.
A. central issue

B. global issue
C. minor issue
D. racial issue

30. Learning is an _____ _____ that continues throughout your life.
    A. actual process
    B. entire process
    C. ongoing process
    D. aging process
31. No _____ _____ has been found to support this claim.
    A. research instrument
    B. research ethics
    C. research interest
    D. research evidence
32. A _____ _____ of these courses is that they are both intended to discuss different social problems related to parenthood.
    A. common feature
    B. physical feature
    C. natural feature
    D. semantic feature
33. As a general rule, new students take a placement test to determine which is the most _____ _____ for them.
    A. appropriate level
    B. elementary level
    C. individual level
    D. proficient level
34. In _____ _____, students are involved in doing activities outside of the classroom instead of learning from textbooks.
    A. disciplinary learning
    B. experiential learning
    C. sequential learning
    D. theoretical learning
35. The new policy only affected _____ _____ companies who started operating from this year.
    A. newly acquired
    B. newly formed
    C. newly issued
    D. newly planted
36. _____ _____ should be given to issues of health and safety.
    A. Careful consideration
    B. Domestic consideration
    C. Favourable consideration
    D. Underlying consideration

37. The two studies follow the same design. A(n) _____ _____ between them is in the activities.
    A. essential difference
    B. fundamental difference
    C. regional difference
    D. systematic difference
38. Both players made a _____ _____ to the team success.
    A. major challenge
    B. major depression
    C. major purpose
    D. major contribution
39. Sales _____ a _____ during the winter months before going down in the spring.
    A. reached a point
    B. reached an end
    C. reached a level
    D. reached a peak
40. Using games and songs may _____ _____ of foreign languages.
    A. enjoy learning
    B. assess learning
    C. enhance learning
    D. generalise learning
41. A record on weather over the past five years is an example of _____ _____.
    A. factual information
    B. medical information
    C. personal information
    D. financial information
42. Greater contact between the two groups should lead to a more _____ _____.
    A. adequate understanding
    B. complete understanding
    C. mutual understanding
    D. original understanding
43. The first period of a project is called the _____ _____.
    A. initial plan
    B. initial data
    C. initial goal
    D. initial phase
44. The _____ _____ means the major purpose for which a building or equipment is intended.
    A. practical function
    B. useful function

C. specific function
D. primary function

45. The sky looks dark. There is a _____ _____ that it will rain tonight.
A. high ability
B. high degree
C. high frequency
D. high probability

46. The use of sounds and words to express yourself is called _____ _____.
A. open communication
B. verbal communication
C. written communication
D. visual communication

47. There are a(n) _____ _____ of stars in the sky.
A. average number
B. small number
C. total number
D. vast number

48. _____ _____ is the changing of the sun's energy into heat and electricity.
A. Absolute power
B. Global power
C. Relative power
D. Solar power

49. As an educator and researcher, she is _____ _____ of individual differences.
A. dimly aware
B. hardly aware
C. keenly aware
D. newly aware

50. This calculator helps to _____ a _____ in just one second.
A. analyse a result
B. interpret a result
C. obtain a result
D. publish a result

51. It is important to _____ _____ in the workplace. It can help to build a good working relationship when people can talk and understand what is said to them.
A. communicate effectively
B. function effectively
C. operate effectively
D. compete effectively

52. A(n) _____ _____ includes people of similar abilities.
A. heterogeneous group
B. indigenous group
C. religious group
D. homogeneous group

53. Their _____ _____ has been extremely influential in the field of science.
    A. clerical work
    B. pastoral work
    C. seminal work
    D. tedious work
54. An article is considered _____ _____ if it reports the researchers' own work.
    A. empirical research
    B. longitudinal research
    C. original research
    D. published research
55. Health care, housing assistance, and childcare assistance are examples of _____ _____.
    A. public record
    B. public opinion
    C. public relation
    D. public welfare
56. Eight kilometres is _____ _____ to five miles. Specifically, five miles is 8.04672 kilometres.
    A. exactly equal
    B. fully equal
    C. roughly equal
    D. truly equal
57. The right of a citizen to travel within a country, and to leave and return to that country is called _____ _____.
    A. free movement
    B. mass movement
    C. national movement
    D. social movement
58. The paper just needs _____ _____ in spelling to get printed.
    A. major changes
    B. gradual changes
    C. massive changes
    D. minor changes
59. _____ _____is the highest level of leadership in an organization.
    A. Strategic management
    B. Internal management
    C. Proper management
    D. Senior management
60. The law had a _____ _____ on house prices. The prices eventually went up at least three times.
    A. relative effect
    B. dramatic effect
    C. potential effect
    D. random effect

Appendix D

The Academic Collocation Recall Test

This is the Recall Test of academic collocations. It tests your ability to produce common written phrases in academic contexts such as in university textbooks or in research papers. There are 60 items in this test.

Please fill in a suitable **academic phrase** in each of the following blanks. **Two initial letters** of each word have been provided. The **meaning of each phrase** is provided in the brackets at the end of each test item. You are not allowed to use the words that have been given in the sentence.

Example:

0. *The region in a city is called an* ***ur*** *ban*_____ ***ar*** *ea*_____. *(city zone)*

1. The two approaches have a **si_____ di_____**. One uses interviews while the other does not. (large variation)
2. Parents also should encourage their children to get involved in a **ph_____ ac_____** such as swimming or running. (bodily movement or exercise)
3. One of the tips to stay healthy is eating a **wi_____ ra_____** of fruits and vegetables. (covering many types)
4. An organization was set up to **ad_____ the is_____** of traffic jams. (think about and begin to deal with a problem)
5. People may not want to share their personal information including age, salary and **re_____ be_____**. (faith in a religion)
6. A set of ideas or plans that a country uses as a basis for making decisions is called **go_____ po_____**. (national plan)

7. An unwritten rule of manners that are considered acceptable in a group or society is called a **so_____ no_____**. (community standard)
8. Customer spending was the most important **dr_____ fo_____** behind the economic growth in the summer. (main reason)
9. The two conditions cannot happen at the same time. They are **mu_____ ex_____**. (not being in agreement)
10. Everything in that event was **we_____ do_____**, including the location, the guest list and the speeches. (clearly written or noted)
11. The interviews were given to a **ra_____ sa_____** of students. Any student was able to join if they wanted. (selection of parts of a unit based on chance)
12. The **ul_____ go_____** of this program is to prevent further damage to the sea. (key purpose)
13. The math teacher **of_____ an op_____** for his students who missed the final exam to take the test again. (gave a chance)
14. Research has found that students' grades and their relationship with parents are **po_____ as_____**. (connected in the same direction)
15. The event welcomed a **di_____ gr_____** of guests from over 30 countries. (different classes)
16. Research results show that (a) **hi_____ pe_____** of smokers also tend to be drinkers. (large number)
17. The university is highly popular and runs various programs in a wide number of **ac_____ di_____**. (fields of study)
18. Science is an **es_____ co_____** of society. (key part)

19. There is more work that needs to be done to **fu_____ un_____** the effect of this change. (totally get the meaning of)
20. Many **ex_____ fa_____** can affect your business such as economic, political and social environment of the locations where the company operates. (outside influence)
21. The library and the union building are located in **cl_____ pr_____**. (short distance to each other)
22. The difference in wages **re_____ co_____**, neither increasing nor reducing over time. (was unchanged)
23. An increase in the use of private vehicles **po_____ a th_____** to the environment. (causes a risk)
24. It will be harder to get the job without **pr_____ ex_____** in the field. (having done something before)
25. The **in_____ va_____** of a dollar bill is only that of a piece of paper. (worth in itself)
26. The government tried to prevent the **wi_____ us_____** of the new drug because of its side effects. (broad application)
27. Water pollution leads to a **hi_____ in_____** of disease. (large degree)
28. Short sentences should be used when **de_____ a pr_____**, such as the instructions on how to build a model. (explaining the way to do something)
29. Air pollution has become the **ce_____ is_____** that affects the development of tourism in this city. (main problem)
30. Learning is an **on_____ pr_____** that continues throughout your life. (continuing exercise)
31. No **re_____ ev_____** has been found to support this claim. (study results)

32. A **co_____ fe_____** of these courses is that they are both intended to discuss different social problems related to parenthood. (similarity)
33. As a general rule, new students take a placement test to determine the most **ap_____ le_____** for them. (suitable degree)
34. In **ex_____ le_____**, students are involved in doing activities outside of the classroom instead of learning from textbooks. (practical training)
35. The new policy only affected **ne_____ fo_____** companies who started operating from this year. (recently created)
36. **Ca_____ co_____** should be given to issues of health and safety. (serious attention)
37. The two studies follow the same design. A **fu_____ di_____** between them is in the activities. (important dissimilarity)
38. Both players made a **ma_____ co_____** to the team success. (important role)
39. Sales **re_____ a pe_____** during the winter months before going down in the spring. (get to the highest point)
40. Using games and songs may **en_____ le_____** of foreign languages. (improve education)
41. The record on weather over the past five years is an example of **fa_____ in_____**. (actual news)
42. Greater contact between the two groups should lead to more **mu_____ un_____**. (common agreement)
43. The first period of a project is called the **in_____ ph_____**. (early stage)
44. The **pr_____ fu_____** means the major purpose for which a building or equipment is intended. (main role)
45. The sky looks dark. There is a **hi_____ pr_____** that it will rain tonight. (good chance)
46. The use of sounds and words to express yourself is called **ve_____ co_____**. (speaking)
47. There are a **va_____ nu_____** of stars in the sky. (so many)

48. **So_____ po_____** is the changing of the sun's energy into heat and electricity. (sun’s energy)
49. As an educator and researcher, she is **ke_____ aw_____** of individual differences. (very conscious)
50. This calculator helps to **ob_____ a re_____** in just one second. (get a finding)
51. It is important to **co_____ ef_____** in the workplace. It can help to build a good working relationship when people can talk and understand what is said to them. (talk in a good way)
52. A **ho_____ gr_____** includes people of shared abilities. (a collection of similar people)
53. Their **se_____ wo_____** has been extremely influential in the field of science. (important research)
54. An article is considered **or_____ re_____** if it reports the researchers’ own work. (a study showing a totally new idea)
55. Health care, housing assistance, and childcare assistance are examples of **pu_____ we_____**. (well-being of a society)
56. Eight kilometres is **ro_____ eq_____** to five miles. Specifically, five miles is 8.04672 kilometres. (more or less the same)
57. The right of a citizen to travel within a country, and to leave and return to that country is called **fr_____ mo_____**. (right to come and go)
58. The paper just needs **mi_____ ch_____** in spelling to get printed. (small difference)
59. **Se_____ ma_____** is the highest level of leadership in an organization. (top leadership)
60. The law had a **dr_____ ef_____** on house prices. The prices eventually went up at least three times. (a great change which is a result of an action)

Appendix E

Rasch Item Measures and Fit Statistics of the AC Recognition Test

Items	Measure (logit)	Infit MNSQ	Infit ZSTD	Point-measure correlation
1	-0.85	1.02	0.22	0.39
2	-2.20	0.86	-0.68	0.30
3	-1.62	0.93	-0.44	0.34
4	-1.42	0.86	-1.21	0.40
5	-1.46	0.91	-0.70	0.37
6	-1.50	0.98	-0.12	0.31
7	-0.82	0.88	-1.33	0.46
8	1.11	1.05	0.68	0.58
9	-0.13	1.06	0.78	0.47
10	0.92	1.31	4.17	0.46
11	0.14	0.80	-3.12	0.60
12	-0.91	0.98	-0.20	0.41
13	-1.62	0.83	-1.27	0.38
14	1.00	1.05	0.79	0.57
15	-0.23	1.06	0.84	0.45
16	-0.13	1.06	0.84	0.44
17	0.82	1.00	0.01	0.58
18	0.70	1.27	3.70	0.46
19	0.76	1.24	3.35	0.47
20	-0.16	1.01	0.12	0.47
21	0.49	1.13	1.84	0.50
22	0.22	1.00	0.09	0.52
23	0.37	0.89	-1.71	0.59
24	-0.42	0.85	-1.99	0.52
25	1.19	1.23	3.11	0.51
26	0.11	1.06	0.92	0.48
27	1.00	1.13	1.87	0.54
28	-0.13	0.88	-1.77	0.54
29	0.33	1.05	0.77	0.51
30	-0.45	0.93	-0.93	0.48
31	-0.35	0.80	-2.86	0.55
32	0.35	0.98	-0.25	0.53
33	0.11	1.10	1.39	0.47
34	-0.13	1.03	0.37	0.47
35	0.45	0.93	-1.02	0.58
36	-0.09	0.98	-0.23	0.50

Items	Measure (logit)	Infit MNSQ	Infit ZSTD	Point-measure correlation
37	1.29	1.26	3.42	0.50
38	0.16	0.76	-3.82	0.62
39	0.49	0.81	-2.90	0.63
40	-0.63	0.94	-0.70	0.45
41	-0.60	0.88	-1.50	0.47
42	-0.06	0.97	-0.34	0.50
43	0.29	1.01	0.16	0.53
44	0.96	1.20	2.75	0.51
45	0.49	0.88	-1.77	0.61
46	-1.27	0.94	-0.48	0.39
47	-0.11	0.84	-2.34	0.56
48	-1.50	0.85	-1.19	0.40
49	0.16	0.98	-0.25	0.53
50	0.05	1.00	-0.03	0.51
51	-0.30	1.06	0.77	0.43
52	0.78	1.01	0.14	0.57
53	0.68	1.53	6.72	0.34
54	0.56	1.08	1.19	0.53
55	0.14	0.81	-2.95	0.59
56	0.53	0.86	-2.20	0.61
57	0.49	1.02	0.27	0.55
58	0.37	0.87	-1.91	0.59
59	0.56	0.86	-2.11	0.61
60	1.05	1.09	1.30	0.56

Appendix F

Rasch Item Measures and Fit Statistics of the AC Recall Test

Items	Measure (logit)	Infit MNSQ	Infit ZSTD	Point-measure correlation
1	0.54	1.00	0.08	0.53
2	-4.72	1.10	0.51	0.39
3	-2.27	1.09	1.09	0.49
4	0.49	1.00	-0.01	0.53
5	-0.92	1.09	1.44	0.52
6	-1.71	1.28	3.59	0.41
7	-1.29	1.18	2.60	0.48
8	0.94	0.92	-1.04	0.54
9	0.84	1.15	2.01	0.44
10	2.18	0.98	-0.15	0.44
11	0.08	0.79	-3.56	0.64
12	-1.75	0.94	-0.76	0.57
13	-3.39	0.99	-0.06	0.50
14	1.93	0.76	-2.68	0.55
15	0.06	1.17	2.65	0.47
16	-1.71	1.18	2.39	0.48
17	0.13	1.11	1.75	0.50
18	-0.17	1.25	3.76	0.45
19	0.17	1.07	1.14	0.51
20	-1.03	1.08	1.22	0.53
21	-0.01	0.98	-0.32	0.56
22	0.08	0.83	-2.88	0.63
23	-0.38	0.81	-3.23	0.64
24	-0.58	1.22	3.34	0.46
25	1.63	1.06	0.75	0.44
26	-1.29	0.86	-2.14	0.62
27	4.27	0.92	-0.19	0.28
28	0.78	0.95	-0.67	0.55
29	-0.83	1.09	1.48	0.53
30	-1.07	0.90	-1.64	0.60
31	0.90	1.07	1.02	0.48
32	0.54	1.00	-0.04	0.54
33	-0.49	1.20	3.15	0.47
34	2.77	1.17	1.23	0.30
35	0.37	0.80	-3.19	0.63
36	1.11	0.98	-0.24	0.51

Items	Measure (logit)	Infit MNSQ	Infit ZSTD	Point-measure correlation
37	0.43	0.90	-1.55	0.58
38	0.64	0.80	-3.07	0.62
39	-0.72	0.83	-2.97	0.64
40	0.66	1.26	3.52	0.40
41	0.15	1.00	0.00	0.55
42	-0.87	0.83	-2.95	0.65
43	-0.63	1.01	0.14	0.56
44	-0.38	1.17	2.60	0.49
45	-0.31	0.93	-1.14	0.58
46	-1.43	1.06	0.94	0.53
47	1.25	1.00	0.06	0.50
48	-2.93	1.01	0.10	0.50
49	1.29	0.94	-0.76	0.53
50	-0.46	0.98	-0.29	0.57
51	-0.42	0.94	-1.05	0.59
52	1.29	0.98	-0.26	0.50
53	3.23	0.72	-1.73	0.45
54	0.56	0.92	-1.16	0.56
55	0.60	1.09	1.33	0.49
56	0.12	0.73	-4.62	0.67
57	1.18	1.27	3.25	0.38
58	-0.40	0.98	-0.37	0.57
59	0.36	0.83	-2.79	0.62
60	0.54	0.99	-0.18	0.54

Appendix G

Frequency of Academic Collocations on COCA Academic and Number of Correct Answers for Each Test Item

		Number of correct answers	
Test item	Frequency	AC Recognition Test	AC Recall Test
1	7220	292	117
2	5957	325	325
3	3020	314	264
4	2128	309	120
5	1065	310	198
6	819	311	239
7	704	291	218
8	587	204	97
9	563	264	102
10	550	214	48
11	528	252	142
12	515	294	241
13	495	314	301
14	494	210	56
15	470	268	143
16	434	264	239
17	426	219	139
18	383	225	156
19	379	222	137
20	355	265	204
21	353	235	147
22	334	248	142
23	314	241	168
24	313	276	179
25	297	200	67
26	296	253	218
27	290	210	10
28	284	264	105
29	272	243	193
30	262	277	206
31	246	273	99
32	236	242	117
33	230	253	174
34	226	264	32
35	224	237	126
36	213	262	89

Test item	Frequency	Number of correct answers	
		AC Recognition Test	AC Recall Test
37	205	195	123
38	201	251	112
39	197	235	187
40	189	284	111
41	189	283	138
42	180	261	195
43	169	245	182
44	167	212	168
45	167	235	164
46	160	305	225
47	157	263	83
48	155	311	288
49	152	251	81
50	150	256	172
51	142	271	170
52	138	221	81
53	137	226	23
54	134	232	116
55	127	252	114
56	126	233	140
57	117	235	86
58	114	241	169
59	113	232	127
60	112	207	117

Appendix H

Top 500 ACL and AECL Items

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
1	mental health	7122
2	higher education	6827
3	physical activity	5143
4	statistically significant	4763
5	foreign policy	4359
6	climate change	3755
7	professional development	3456
8	federal government	3336
9	future research	3113
10	wide range	2821
11	significant difference	2749
12	economic growth	2479
13	academic achievement	2472
14	national security	2449
15	public policy	2370
16	significantly higher	2345
17	civil society	2293
18	previous research	2170
19	critical thinking	2127
20	environmental protection	2126
21	private sector	2026
22	high level	1710
23	vast majority	1575
24	popular culture	1567
25	academic performance	1562
26	international community	1515
27	further research	1492
28	dependent variable	1443
29	primary care	1414
30	socioeconomic status	1347
31	have access	1346
32	sexual orientation	1346
33	closely related	1238
34	data set	1195
35	due process	1180
36	informed consent	1180
37	natural history	1144
38	domestic violence	1140
39	sexual abuse	1135

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
1	mental health	7122
2	present study	6483
3	university press	5761
4	physical activity	5143
5	statistically significant	4763
6	state university	4550
7	climate change	3755
8	data collection	3501
9	professional development	3456
10	drug use	3332
11	future research	3113
12	current study	3109
13	wide range	2821
14	control group	2790
15	short term	2757
16	significant difference	2749
17	general education	2648
18	social support	2575
19	large scale	2553
20	economic growth	2479
21	association between	2446
22	economic development	2412
23	public policy	2370
24	significantly higher	2345
25	civil society	2293
26	subject matter	2210
27	important role	2192
28	previous research	2170
29	case study	2159
30	total number	2139
31	data analysis	2133
32	critical thinking	2127
33	environmental protection	2126
34	private sector	2026
35	grade level	1962
36	international law	1944
37	low income	1789
38	English language	1730
39	task force	1721

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
40	financial support	1134
41	commonly used	1128
42	learning environment	1128
43	integral part	1115
44	ethnic group	1100
45	statistical analysis	1092
46	national identity	1074
47	intellectual property	1061
48	learning process	1050
49	widely used	1047
50	mental illness	1039
51	randomly selected	1009
52	relatively low	1009
53	ethnic identity	1007
54	local government	1003
55	increased risk	961
56	educational research	946
57	political economy	946
58	educational system	937
59	nuclear power	925
60	statistical significance	923
61	high quality	922
62	international journal	891
63	relatively high	885
64	social interaction	883
65	personal communication	871
66	recent study	871
67	academic year	863
68	foreign investment	857
69	public sector	856
70	strongly agree	854
71	empirical evidence	853
72	ethnic minority	849
73	global economy	836
74	longitudinal study	836
75	significant effect	835
76	recent research	833
77	mean score	823
78	strongly disagree	817
79	qualitative research	816
80	broad range	810
81	next generation	798
82	readily available	798

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
40	high level	1710
41	social science	1623
42	standard deviation	1606
43	national association	1576
44	disease control	1568
45	popular culture	1567
46	internal consistency	1555
47	effect size	1550
48	wide variety	1517
49	international community	1515
50	national institute	1510
51	factor analysis	1503
52	human being	1484
53	high quality	1481
54	significantly different	1463
55	social worker	1455
56	dependent variable	1443
57	primary care	1414
58	long term	1382
59	natural gas	1371
60	relatively small	1366
61	socioeconomic status	1347
62	have access	1346
63	conflict between	1334
64	main effect	1316
65	mean age	1304
66	age group	1301
67	small group	1296
68	regression analysis	1286
69	time period	1283
70	research center	1238
71	general population	1236
72	general public	1234
73	human nature	1234
74	population growth	1202
75	data set	1195
76	higher level	1191
77	visual impairment	1185
78	informed consent	1180
79	focus group	1172
80	research question	1171
81	land use	1150
82	domestic violence	1140

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
83	provide information	795
84	pilot study	793
85	sexual intercourse	792
86	public domain	772
87	central government	770
88	randomly assigned	766
89	academic success	761
90	legal system	749
91	natural world	749
92	scientific research	734
93	independent variable	724
94	significant role	720
95	technical assistance	720
96	additional information	718
97	conceptual framework	711
98	annual meeting	704
99	equally important	704
100	first author	695
101	economic crisis	688
102	cultural diversity	687
103	military service	680
104	current research	677
105	directly related	677
106	secondary education	677
107	prior knowledge	675
108	physical health	672
109	cultural identity	671
110	special issue	669
111	risk assessment	658
112	social context	658
113	local community	656
114	renewable energy	656
115	theoretical framework	654
116	political culture	650
117	key role	649
118	economic policy	647
119	solar system	640
120	central role	633
121	empirical research	630
122	relatively few	629
123	natural resource	628
124	social status	627
125	indigenous people	623

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
83	financial support	1134
84	point scale	1128
85	integral part	1115
86	everyday life	1113
87	high degree	1108
88	ethnic group	1100
89	human life	1099
90	statistical analysis	1092
91	intellectual property	1061
92	human development	1060
93	learning process	1050
94	significantly lower	1041
95	mental illness	1039
96	global warming	1019
97	labor force	1019
98	relatively low	1009
99	response rate	1005
100	labor market	1002
101	important part	992
102	sexual activity	988
103	information technology	969
104	increased risk	961
105	social change	957
106	knowledge base	956
107	present day	942
108	social life	942
109	international trade	936
110	multiple regression	935
111	literature review	924
112	statistical significance	923
113	growth rate	916
114	making process	909
115	international journal	891
116	relatively high	885
117	social interaction	883
118	health organization	882
119	human health	880
120	research council	876
121	recent study	871
122	public sector	856
123	strongly agree	854
124	empirical evidence	853
125	risk factor	852

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
126	major role	623
127	cultural heritage	619
128	positive relationship	618
129	international conference	612
130	political party	603
131	relatively little	598
132	emotional support	597
133	structural adjustment	597
134	negative impact	594
135	significant relationship	592
136	social welfare	589
137	differ significantly	576
138	social structure	575
139	personal experience	574
140	positive effect	570
141	historical context	569
142	well established	569
143	active role	565
144	national survey	565
145	full range	564
146	national conference	564
147	widely accepted	563
148	positive impact	562
149	conflict resolution	558
150	national interest	557
151	survey data	557
152	annual report	556
153	positively correlated	554
154	economic activity	548
155	further study	548
156	social policy	544
157	career development	543
158	frequently used	540
159	military force	535
160	mutually exclusive	533
161	religious freedom	533
162	significant impact	533
163	third party	533
164	missing data	530
165	daily living	529
166	further investigation	528
167	well documented	527
168	critical role	526

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
126	public interest	851
127	ethnic minority	849
128	common sense	842
129	particularly important	839
130	global economy	836
131	longitudinal study	836
132	significant effect	835
133	recent research	833
134	research project	832
135	new technology	828
136	mean score	823
137	long history	822
138	important factor	817
139	strongly disagree	817
140	qualitative research	816
141	broad range	810
142	new information	802
143	readily available	798
144	justice system	795
145	provide information	795
146	pilot study	793
147	research team	791
148	experimental group	790
149	public domain	772
150	central government	770
151	small scale	767
152	have difficulty	763
153	working group	756
154	developing world	749
155	natural world	749
156	learning experience	743
157	review board	741
158	science foundation	735
159	scientific research	734
160	working memory	729
161	census bureau	724
162	independent variable	724
163	social behavior	722
164	treatment group	721
165	significant role	720
166	technical assistance	720
167	additional information	718
168	national level	718

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
169	low level	525
170	natural environment	522
171	significantly correlated	517
172	economic reform	516
173	presidential election	516
174	information processing	515
175	significant increase	514
176	collective action	511
177	public sphere	511
178	liberal democracy	510
179	environmental degradation	507
180	qualitative data	506
181	positive correlation	505
182	metropolitan area	503
183	strategic planning	503
184	provide evidence	496
185	environmental policy	492
186	collect data	487
187	immediately following	486
188	political participation	486
189	primary source	486
190	significant number	485
191	become involved	482
192	economic system	482
193	natural law	478
194	cultural context	476
195	little research	474
196	national government	473
197	racial discrimination	470
198	slightly higher	468
199	experimental design	467
200	clearly defined	466
201	positively associated	464
202	increasingly important	463
203	possible explanation	462
204	slightly different	462
205	conventional wisdom	460
206	negative effect	460
207	military power	458
208	positive attitude	457
209	focal point	455
210	political philosophy	455
211	next decade	453

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
169	nervous system	718
170	exchange rate	717
171	rating scale	716
172	social class	713
173	information system	712
174	conceptual framework	711
175	research institute	710
176	equally important	704
177	economic crisis	688
178	cultural diversity	687
179	local level	687
180	content analysis	685
181	current research	677
182	prior knowledge	675
183	social context	658
184	additional research	657
185	especially important	655
186	following year	655
187	theoretical framework	654
188	key role	649
189	economic policy	647
190	marital status	646
191	natural selection	646
192	foreign exchange	643
193	family planning	641
194	short term	641
195	resource management	639
196	increasing number	637
197	research design	635
198	central role	633
199	limited number	632
200	empirical research	630
201	significantly greater	629
202	natural resource	628
203	social status	627
204	major role	623
205	human behavior	621
206	prior research	620
207	alcohol consumption	618
208	positive relationship	618
209	clinical practice	617
210	regression model	617
211	life expectancy	615

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
212	significant interaction	453
213	crucial role	451
214	random sample	451
215	highly correlated	450
216	primary purpose	450
217	little evidence	448
218	provide support	446
219	actively involved	441
220	cognitive development	441
221	ultimate goal	438
222	scientific community	437
223	significant change	436
224	economic status	435
225	annual conference	428
226	economic power	427
227	past research	425
228	social environment	425
229	dominant culture	424
230	social organization	424
231	media coverage	422
232	further evidence	417
233	overwhelming majority	416
234	subject area	416
235	well aware	414
236	creative thinking	413
237	government policy	413
238	primary focus	413
239	human activity	412
240	previous study	412
241	driving force	411
242	social identity	410
243	detailed information	409
244	first phase	407
245	become aware	400
246	major source	399
247	available evidence	397
248	social responsibility	396
249	specific information	396
250	political reform	395
251	first generation	394
252	financial assistance	393
253	democratic society	392
254	scientific evidence	391

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
212	care system	614
213	demographic information	612
214	international conference	612
215	new knowledge	612
216	content area	610
217	make use	609
218	advisory committee	607
219	international system	606
220	extended family	605
221	highest level	605
222	previous year	599
223	relatively little	598
224	emotional support	597
225	data suggest	596
226	next section	596
227	decision making	594
228	human body	594
229	negative impact	594
230	significant relationship	592
231	low cost	591
232	modern world	589
233	peer group	586
234	social network	585
235	education system	583
236	skill development	579
237	social structure	575
238	personal experience	574
239	positive effect	570
240	historical context	569
241	active role	565
242	mental retardation	565
243	national survey	565
244	full range	564
245	human experience	564
246	positive impact	562
247	regular basis	558
248	national interest	557
249	survey data	557
250	total population	557
251	good example	555
252	multiple choice	553
253	human capital	551
254	important aspect	551

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
255	social movement	389
256	active participation	386
257	lifelong learning	386
258	standard error	385
259	economic integration	383
260	positive feedback	382
261	environmental impact	380
262	assessment process	378
263	generally accepted	378
264	medical treatment	378
265	social isolation	378
266	cultural history	377
267	raw data	377
268	social integration	377
269	useful information	377
270	high priority	376
271	high rate	376
272	negatively correlated	375
273	short period	374
274	public debate	371
275	further analysis	370
276	political stability	366
277	significant correlation	365
278	greater emphasis	364
279	previously described	364
280	significant amount	363
281	radically different	362
282	gain access	360
283	artificial intelligence	359
284	creative process	359
285	significant improvement	359
286	currently available	357
287	highly significant	354
288	numerous studies	353
289	security policy	353
290	relevant information	352
291	final analysis	350
292	major problem	350
293	national policy	350
294	oral history	350
295	environmental change	348
296	available data	347
297	historical perspective	345

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
255	odds ratio	551
256	public support	549
257	economic activity	548
258	study period	548
259	control condition	546
260	birth control	545
261	significant predictor	545
262	social policy	544
263	effective way	540
264	life cycle	539
265	formal education	538
266	special interest	538
267	total score	535
268	mutually exclusive	533
269	significant impact	533
270	missing data	530
271	critical role	526
272	low level	525
273	research unit	525
274	computer science	524
275	comparison group	522
276	natural environment	522
277	particular interest	520
278	daily basis	519
279	continuing education	518
280	information processing	515
281	significant increase	514
282	management system	512
283	public sphere	511
284	new system	509
285	primary school	509
286	social capital	507
287	work environment	507
288	qualitative data	506
289	positive correlation	505
290	total sample	502
291	common ground	498
292	background information	497
293	planning process	496
294	provide evidence	496
295	scientific knowledge	495
296	skill level	495
297	demographic data	492

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
298	organizational structure	345
299	relatively stable	344
300	draw attention	342
301	closely associated	341
302	minority group	341
303	technological change	340
304	previous work	339
305	small percentage	339
306	strong evidence	339
307	physical environment	338
308	high percentage	337
309	historical record	336
310	military action	336
311	close proximity	335
312	entirely different	335
313	take responsibility	335
314	empirical data	333
315	closer look	332
316	political agenda	331
317	physical appearance	330
318	qualitative study	330
319	direct contact	329
320	rapidly changing	329
321	large proportion	328
322	get involved	326
323	internal control	325
324	open access	325
325	public awareness	325
326	armed conflict	324
327	religious belief	323
328	valuable information	322
329	empirical support	321
330	closely linked	320
331	major concern	320
332	negative correlation	320
333	naturally occurring	319
334	minimum wage	318
335	background knowledge	317
336	quantitative data	317
337	professional practice	314
338	significant contribution	314
339	younger generation	314
340	potential impact	313

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
298	goal setting	492
299	research literature	492
300	current state	490
301	collect data	487
302	social history	487
303	primary source	486
304	significant number	485
305	lesser extent	484
306	rapid growth	483
307	economic system	482
308	extremely important	482
309	age range	481
310	intellectual disability	481
311	see appendix	480
312	human history	479
313	cultural context	476
314	little research	474
315	study area	474
316	experimental design	467
317	increasingly important	463
318	possible explanation	462
319	representative sample	462
320	course content	461
321	important component	461
322	conventional wisdom	460
323	negative effect	460
324	study suggest	460
325	peer review	458
326	positive attitude	457
327	focal point	455
328	primary goal	455
329	relative importance	455
330	new approach	454
331	social learning	454
332	significant interaction	453
333	food production	452
334	crucial role	451
335	random sample	451
336	primary purpose	450
337	multivariate analysis	447
338	provide support	446
339	family history	445
340	blood flow	443

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
341	further information	312
342	collective identity	311
343	comparative analysis	311
344	wide array	311
345	major factor	310
346	close relationship	309
347	earlier version	309
348	gather information	309
349	legal status	309
350	provide insight	309
351	equal opportunity	308
352	fully understand	307
353	entirely new	305
354	general agreement	304
355	political arena	303
356	widely recognized	303
357	social mobility	301
358	public discourse	300
359	key element	299
360	particularly useful	298
361	basic research	297
362	personal responsibility	297
363	emotional intelligence	296
364	significant portion	296
365	further development	295
366	additional support	294
367	great majority	294
368	key component	293
369	low income	293
370	natural science	293
371	scientific method	293
372	strongly associated	293
373	collaborative learning	292
374	frequently cited	292
375	anecdotal evidence	291
376	equal access	291
377	political authority	291
378	ruling party	291
379	sharp contrast	291
380	cultural change	290
381	modern society	290
382	purchasing power	290
383	political context	289

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
341	cognitive development	441
342	ultimate goal	438
343	scientific community	437
344	significant change	436
345	data indicate	435
346	economic status	435
347	month period	434
348	mortality rate	432
349	relatively new	432
350	long period	426
351	past research	425
352	social environment	425
353	dominant culture	424
354	social organization	424
355	quality control	422
356	population density	418
357	tobacco use	418
358	subject area	416
359	expert system	414
360	government policy	413
361	primary focus	413
362	human activity	412
363	previous study	412
364	design process	409
365	detailed information	409
366	research laboratory	409
367	given time	408
368	group membership	407
369	longer term	402
370	large extent	401
371	attitude towards	399
372	major source	399
373	native language	399
374	survey instrument	398
375	available evidence	397
376	last century	396
377	social responsibility	396
378	specific information	396
379	university faculty	396
380	deeper understanding	395
381	relatively large	394
382	advisory board	393
383	financial assistance	393

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
384	significant factor	289
385	vary widely	289
386	previously mentioned	288
387	test score	288
388	gender equality	287
389	specifically designed	286
390	extended period	285
391	significantly reduced	285
392	easy access	282
393	widely available	282
394	annual review	281
395	economic analysis	281
396	highly effective	281
397	key factor	281
398	far removed	280
399	political organization	280
400	economic theory	277
401	high proportion	277
402	widespread use	277
403	direct observation	276
404	electronic media	276
405	field research	276
406	rapidly growing	275
407	critical analysis	274
408	effective communication	274
409	directly involved	273
410	fundamentally different	273
411	personal information	273
412	primary concern	273
413	widely known	273
414	educational policy	272
415	become available	271
416	final section	271
417	physical world	270
418	average score	268
419	highly valued	267
420	vital role	267
421	deeply rooted	265
422	existing research	265
423	generally considered	265
424	internet access	265
425	public administration	265
426	high incidence	264

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
384	democratic society	392
385	percent increase	391
386	science fiction	391
387	scientific evidence	391
388	control system	389
389	null hypothesis	389
390	functional analysis	387
391	lower level	387
392	active participation	386
393	important issue	386
394	standard error	385
395	higher percentage	384
396	joint venture	383
397	positive feedback	382
398	environmental impact	380
399	data show	379
400	education level	379
401	assessment process	378
402	general practice	378
403	social isolation	378
404	raw data	377
405	social control	377
406	social integration	377
407	useful information	377
408	high rate	376
409	delivery system	374
410	short period	374
411	week period	373
412	risk management	372
413	effective means	371
414	family structure	370
415	educational level	369
416	selection process	368
417	learning theory	366
418	commercially available	365
419	relatively short	365
420	significant correlation	365
421	western civilization	365
422	greater emphasis	364
423	interest group	364
424	significant amount	363
425	life history	362
426	radically different	362

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
427	comparative study	261
428	large percentage	261
429	current status	260
430	political climate	260
431	qualitative analysis	260
432	relatively simple	260
433	relatively recent	258
434	earlier work	257
435	intrinsic value	257
436	previous experience	257
437	share information	257
438	accurate information	255
439	experimental research	255
440	increasingly difficult	255
441	modified version	255
442	socially constructed	255
443	detailed analysis	254
444	large majority	254
445	strongly influenced	254
446	substantial number	254
447	high value	253
448	adversely affect	252
449	nuclear war	252
450	national average	251
451	welfare reform	251
452	human society	249
453	nuclear energy	249
454	stark contrast	249
455	nuclear weapon	248
456	similar pattern	248
457	economic success	247
458	virtually impossible	247
459	political activity	246
460	considerable amount	245
461	atomic energy	244
462	clear evidence	244
463	government intervention	244
464	peace treaty	243
465	successful implementation	243
466	fully developed	242
467	main source	242
468	general theory	241
469	particularly relevant	241

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
427	individual level	361
428	secondary level	361
429	gain access	360
430	creative process	359
431	significant improvement	359
432	gold standard	358
433	sea level	358
434	agricultural production	357
435	currently available	357
436	industrial revolution	357
437	content validity	355
438	important point	355
439	highly significant	354
440	human mind	354
441	relatively easy	353
442	security policy	353
443	relevant information	352
444	important question	351
445	final analysis	350
446	next century	350
447	population size	350
448	conceptual model	349
449	particular attention	349
450	environmental change	348
451	human resource	348
452	available data	347
453	historical perspective	345
454	organizational structure	345
455	consent form	344
456	high profile	344
457	model fit	344
458	relatively stable	344
459	university school	344
460	western world	344
461	development process	342
462	minority group	341
463	strong support	341
464	wider range	341
465	domestic product	340
466	technological change	340
467	behavior change	339
468	common practice	339
469	previous work	339

Top 500 ACL (Ackermann & Chen, 2013) items		
No.	Item	FRQ
470	primary reason	241
471	conduct research	240
472	large portion	240
473	full potential	239
474	geographic location	239
475	leading role	239
476	national culture	239
477	politically correct	239
478	prominent role	239
479	research evidence	239
480	highly unlikely	238
481	little information	238
482	theoretical model	238
483	unique opportunity	238
484	technical support	237
485	urban development	237
486	economic value	236
487	research methodology	236
488	academic community	235
489	democratic process	235
490	newly created	235
491	seek help	234
492	changing world	232
493	historical development	232
494	modern technology	230
495	positive influence	230
496	primary responsibility	230
497	provide access	230
498	brief history	229
499	pivotal role	229
500	previous section	229

Top 500 Lex AECL (Lei & Liu, 2018) items		
No.	Item	FRQ
470	small percentage	339
471	soil erosion	339
472	greater degree	338
473	physical environment	338
474	high percentage	337
475	important source	337
476	research assistant	337
477	historical record	336
478	close proximity	335
479	market value	335
480	following section	334
481	social group	334
482	empirical data	333
483	medical association	332
484	social system	332
485	qualitative study	330
486	recent history	330
487	direct contact	329
488	social construction	329
489	large proportion	328
490	writing process	328
491	open access	325
492	religious belief	323
493	valuable information	322
494	empirical support	321
495	major concern	320
496	target population	319
497	background knowledge	317
498	human population	317
499	quantitative data	317
500	immune system	316

Appendix I

Overlapping Items Between the ACL and the AECL Ordered by Frequency

1	mental health	436	significant proportion	871	fundamental importance
2	physical activity	437	vested interest	872	greatly enhance
3	statistically significant	438	industrial production	873	prominent feature
4	climate change	439	facilitate development	874	racial difference
5	professional development	440	directly affect	875	scientific theory
6	future research	441	increased awareness	876	slow process
7	wide range	442	technical expertise	877	technical skill
8	significant difference	443	make transition	878	theoretical understanding
9	economic growth	444	brief period	879	use method
10	public policy	445	vital part	880	generally agree
11	significantly higher	446	high probability	881	global network
12	civil society	447	provide data	882	negative view
13	previous research	448	socially desirable	883	normal development
14	critical thinking	449	central issue	884	discuss issue
15	environmental protection	450	direct access	885	alternative view
16	private sector	451	effective method	886	critically evaluate
17	high level	452	manufacturing sector	887	distinctive feature
18	popular culture	453	service sector	888	methodological approach
19	international community	454	complete task	889	numerical data
20	dependent variable	455	highly sensitive	890	overall structure
21	primary care	456	improved performance	891	specific focus
22	socioeconomic status	457	particular area	892	subsequent development
23	have access	458	public access	893	common approach
24	data set	459	brief review	894	current trend
25	informed consent	460	provide service	895	immediate environment
26	play role	461	continued use	896	limited capacity
27	domestic violence	462	greater likelihood	897	new insight
28	financial support	463	historical analysis	898	significant shift
29	integral part	464	potentially dangerous	899	complex structure
30	ethnic group	465	relatively rare	900	experience difficulty
31	statistical analysis	466	relevant literature	901	thought process
32	intellectual property	467	close contact	902	traditional method
33	learning process	468	gather data	903	traditional practice
34	mental illness	469	profound impact	904	briefly discuss
35	relatively low	470	relatively minor	905	direct role
36	increased risk	471	global perspective	906	dramatic effect
37	statistical significance	472	collect information	907	natural process
38	high quality	473	ethnic community	908	tacit knowledge
39	international journal	474	major focus	909	affect outcome
40	relatively high	475	appropriate level	910	provide explanation
41	social interaction	476	natural language	911	use strategy
42	recent study	477	personal relationship	912	accurate description
43	public sector	478	central importance	913	particular aspect
44	strongly agree	479	reduce likelihood	914	possible source
45	empirical evidence	480	basic information	915	related activity
46	ethnic minority	481	future development	916	related problem

47	global economy	482	gain insight	917	basic concept
48	longitudinal study	483	highly dependent	918	emotional reaction
49	significant effect	484	theoretical perspective	919	essential information
50	recent research	485	make distinction	920	specific issue
51	mean score	486	careful analysis	921	biological science
52	strongly disagree	487	public image	922	brief introduction
53	qualitative research	488	dominant group	923	briefly describe
54	broad range	489	linear relationship	924	common source
55	readily available	490	significant degree	925	common usage
56	provide information	491	take precedence	926	enhance performance
57	pilot study	492	competitive market	927	large quantity
58	public domain	493	internal structure	928	specific example
59	central government	494	natural order	929	striking contrast
60	natural world	495	individual behaviour	930	establish relationship
61	scientific research	496	increased number	931	alternative way
62	independent variable	497	integrated approach	932	individual variation
63	significant role	498	local culture	933	precise nature
64	technical assistance	499	sufficient evidence	934	principal source
65	additional information	500	holistic approach	935	highly sophisticated
66	social behaviour	501	systematic approach	936	particularly valuable
67	conceptual framework	502	deep understanding	937	specific question
68	equally important	503	empirical work	938	use approach
69	economic crisis	504	experimental study	939	biological evolution
70	cultural diversity	505	rapid expansion	940	present evidence
71	current research	506	strong emphasis	941	quantitative study
72	prior knowledge	507	central question	942	seek information
73	social context	508	current policy	943	theoretical analysis
74	theoretical framework	509	increasingly common	944	alternative interpretation
75	key role	510	notable exception	945	finite number
76	economic policy	511	remarkably similar	946	gain information
77	central role	512	strongly suggest	947	initial period
78	empirical research	513	traditional approach	948	learning objective
79	natural resource	514	beneficial effect	949	obtain data
80	social status	515	guiding principle	950	offer insight
81	major role	516	profound effect	951	common assumption
82	positive relationship	517	provide assistance	952	current climate
83	international conference	518	complex set	953	numerical value
84	relatively little	519	dominant role	954	previous generation
85	emotional support	520	markedly different	955	traditional form
86	negative impact	521	particular emphasis	956	make impact
87	significant relationship	522	revised version	957	core value
88	various aspects	523	specific type	958	direct communication
89	social structure	524	alternative explanation	959	directly proportional
90	personal experience	525	increased demand	960	salient feature
91	positive effect	526	perform task	961	ethical dilemma
92	historical context	527	provide overview	962	general conclusion
93	active role	528	human interaction	963	increasing proportion
94	national survey	529	encourage development	964	single entity
95	full range	530	effective management	965	contain information
96	positive impact	531	environmental concern	966	main function

97	national interest	532	existing data	967	overall rate
98	survey data	533	increasingly popular	968	strong tendency
99	economic activity	534	information sharing	969	identify problem
100	social policy	535	make available	970	provide indication
101	mutually exclusive	536	pioneering work	971	effective policy
102	significant impact	537	provide alternative	972	give information
103	missing data	538	external environment	973	modified form
104	critical role	539	statistical data	974	comprehensive system
105	low level	540	written communication	975	main task
106	natural environment	541	fundamental problem	976	negative connotation
107	information processing	542	future study	977	negative outcome
108	significant increase	543	high status	978	planning stage
109	public sphere	544	limited information	979	moral dilemma
110	qualitative data	545	particularly effective	980	analytical approach
111	positive correlation	546	quantitative research	981	enormous impact
112	use data	547	critical issue	982	external source
113	provide evidence	548	dramatic change	983	individual item
114	collect data	549	fundamental question	984	intensive study
115	primary source	550	historical background	985	negative value
116	significant number	551	key feature	986	original model
117	economic system	552	significantly increase	987	previous discussion
118	cultural context	553	complex relationship	988	meet requirement
119	little research	554	perceived importance	989	similar result
120	experimental design	555	theoretical basis	990	alternative form
121	increasingly important	556	increased interest	991	general overview
122	address issue	557	potential problem	992	legal requirement
123	possible explanation	558	practical significance	993	social aspect
124	conventional wisdom	559	prove useful	994	subsequent study
125	negative effect	560	technical knowledge	995	substantial difference
126	positive attitude	561	international agreement	996	draw distinction
127	focal point	562	basic structure	997	abstract concept
128	significant interaction	563	necessary information	998	broadly similar
129	crucial role	564	general consensus	999	earlier discussion
130	random sample	565	highly complex	1000	general category
131	primary purpose	566	largely responsible	1001	particularly apparent
132	provide support	567	major theme	1002	previous paragraph
133	cognitive development	568	substantially different	1003	social relationship
134	ultimate goal	569	theoretical work	1004	convey meaning
135	scientific community	570	critical point	1005	ethical problem
136	significant change	571	perceived need	1006	minimum standard
137	economic status	572	reliable data	1007	possible outcome
138	past research	573	reliable information	1008	preceding section
139	social environment	574	report data	1009	standard method
140	dominant culture	575	cultural practice	1010	strong link
141	social organization	576	low priority	1011	certain assumption
142	subject area	577	seminal work	1012	pose challenge
143	government policy	578	deeper level	1013	use technique
144	primary focus	579	economic stability	1014	central core
145	human activity	580	potential risk	1015	clearly important
146	previous study	581	relevant data	1016	considerable importance

147	detailed information	582	complex system	1017	core element
148	major source	583	increased competition	1018	existing structure
149	available evidence	584	plausible explanation	1019	negative consequence
150	social responsibility	585	valuable resource	1020	positive aspect
151	specific information	586	hierarchical structure	1021	related question
152	financial assistance	587	increasing interest	1022	statistical technique
153	democratic society	588	initial phase	1023	external force
154	scientific evidence	589	positive outcome	1024	low probability
155	active participation	590	primary function	1025	obvious difference
156	standard error	591	assume role	1026	central concept
157	positive feedback	592	evolutionary process	1027	ethical issue
158	environmental impact	593	free access	1028	external influence
159	numerous studies	594	negative side	1029	living standard
160	assessment process	595	similar situation	1030	provide context
161	social isolation	596	symbiotic relationship	1031	scarce resource
162	raw data	597	systematic analysis	1032	significant feature
163	social integration	598	traditional view	1033	develop theory
164	useful information	599	brief summary	1034	explore issue
165	high rate	600	common feature	1035	perform function
166	short period	601	increasing pressure	1036	characteristic feature
167	provide opportunity	602	potential source	1037	comprehensive account
168	significant correlation	603	relatively common	1038	main factor
169	greater emphasis	604	urban environment	1039	common characteristic
170	significant amount	605	assume responsibility	1040	distinct group
171	radically different	606	general trend	1041	ethnic difference
172	gain access	607	similar approach	1042	flexible approach
173	creative process	608	vast array	1043	maintain contact
174	significant improvement	609	detailed study	1044	specific aspect
175	currently available	610	infinite number	1045	technical problem
176	highly significant	611	limited range	1046	vast range
177	security policy	612	narrow range	1047	broad agreement
178	relevant information	613	physical science	1048	fundamental assumption
179	increase likelihood	614	readily accessible	1049	separate entity
180	final analysis	615	roughly equal	1050	conduct survey
181	environmental change	616	social inequality	1051	main feature
182	available data	617	specific area	1052	regional variation
183	historical perspective	618	creative work	1053	schematic representation
184	organizational structure	619	examine role	1054	transport system
185	relatively stable	620	direct link	1055	additional problem
186	minority group	621	preliminary data	1056	environmental factor
187	technological change	622	appropriate response	1057	full information
188	previous work	623	central feature	1058	provide coverage
189	small percentage	624	considerable evidence	1059	analytical tool
190	physical environment	625	economic structure	1060	broad category
191	high percentage	626	high profile	1061	key characteristic
192	historical record	627	highly selective	1062	minimum requirement
193	close proximity	628	historical period	1063	conduct analysis
194	empirical data	629	intimate relationship	1064	disclose information
195	qualitative study	630	teaching strategy	1065	general statement
196	direct contact	631	clear distinction	1066	interpret data

197	large proportion	632	entire range	1067	negative aspect
198	open access	633	particularly evident	1068	process data
199	religious belief	634	social activity	1069	recurrent theme
200	valuable information	635	substantial evidence	1070	draw conclusion
201	empirical support	636	dynamic process	1071	correct interpretation
202	major concern	637	initial stage	1072	natural condition
203	background knowledge	638	marked contrast	1073	statistical method
204	quantitative data	639	social function	1074	underlying principle
205	professional practice	640	superior performance	1075	provide source
206	significant contribution	641	free movement	1076	classical theory
207	potential impact	642	historical event	1077	complex pattern
208	comparative analysis	643	increasing demand	1078	fundamental aspect
209	wide array	644	particularly significant	1079	introductory section
210	major factor	645	source material	1080	learning outcome
211	close relationship	646	freely available	1081	overall aim
212	gather information	647	high standard	1082	place emphasis
213	legal status	648	increased pressure	1083	specific function
214	provide insight	649	sufficient condition	1084	develop strategy
215	general agreement	650	acquire knowledge	1085	appropriate form
216	key element	651	desired outcome	1086	ethical principle
217	particularly useful	652	employment opportunity	1087	minimum value
218	basic research	653	make explicit	1088	relevant factor
219	significant portion	654	ongoing debate	1089	research finding
220	additional support	655	professional knowledge	1090	transmit information
221	key component	656	ready access	1091	provide benefit
222	low income	657	senior management	1092	alternative strategy
223	natural science	658	visual perception	1093	current issue
224	scientific method	659	future prospects	1094	give insight
225	anecdotal evidence	660	appropriate treatment	1095	alternative source
226	equal access	661	critical evaluation	1096	small quantity
227	sharp contrast	662	defining characteristic	1097	store information
228	cultural change	663	environmental pollution	1098	provide summary
229	modern society	664	geographic distribution	1099	serve function
230	purchasing power	665	greater flexibility	1100	basic element
231	political context	666	historical study	1101	causal relation
232	significant factor	667	highly desirable	1102	dynamic system
233	gender equality	668	learning strategy	1103	ethical question
234	extended period	669	next phase	1104	experimental method
235	easy access	670	reciprocal relationship	1105	secondary source
236	widely available	671	sufficient information	1106	extract information
237	annual review	672	diverse range	1107	key finding
238	economic analysis	673	increasing importance	1108	main finding
239	highly effective	674	relatively straightforward	1109	normal practice
240	key factor	675	clearly visible	1110	qualitative method
241	economic theory	676	single individual	1111	useful means
242	high proportion	677	central point	1112	alternative solution
243	widespread use	678	fundamental principle	1113	common error
244	direct observation	679	highly relevant	1114	considerable detail
245	critical analysis	680	increasingly sophisticated	1115	specific need
246	effective communication	681	overall level	1116	take role

247	fundamentally different	682	provide care	1117	established practice
248	personal information	683	social background	1118	little significance
249	primary concern	684	crucial point	1119	relevant material
250	final section	685	developmental process	1120	theoretical concept
251	vital role	686	economic benefit	1121	full analysis
252	existing research	687	have potential	1122	major feature
253	comparative study	688	underlying cause	1123	professional activity
254	large percentage	689	extremely useful	1124	public attitude
255	appropriate behaviour	690	global context	1125	single element
256	current status	691	highly likely	1126	technological advance
257	political climate	692	unintended consequence	1127	external factor
258	qualitative analysis	693	conduct study	1128	published material
259	relatively simple	694	common method	1129	quantitative approach
260	relatively recent	695	dynamic nature	1130	random variable
261	intrinsic value	696	fairly common	1131	related factor
262	previous experience	697	major contribution	1132	technical term
263	share information	698	single source	1133	take approach
264	accurate information	699	resolve conflict	1134	accurate record
265	experimental research	700	causal link	1135	industrialized country
266	increasingly difficult	701	considerable interest	1136	industrialized nation
267	modified version	702	combined effect	1137	related area
268	promote development	703	environmental issue	1138	require knowledge
269	detailed analysis	704	key aspect	1139	develop method
270	substantial number	705	literal meaning	1140	cultural institution
271	high value	706	particularly sensitive	1141	deny access
272	adversely affect	707	appropriate way	1142	individual component
273	human society	708	crucial part	1143	normal condition
274	stark contrast	709	fundamental difference	1144	particular feature
275	similar pattern	710	individual difference	1145	meet objective
276	economic success	711	social norm	1146	basic component
277	considerable amount	712	substantial part	1147	basic function
278	successful implementation	713	make contribution	1148	essential function
279	main source	714	basic principle	1149	methodological problem
280	general theory	715	considerable effort	1150	provide material
281	particularly relevant	716	considerable influence	1151	related topic
282	primary reason	717	cultural value	1152	structural feature
283	conduct research	718	increasing awareness	1153	undertake research
284	full potential	719	main theme	1154	make adjustment
285	leading role	720	major shift	1155	comprehensive overview
286	prominent role	721	primary aim	1156	key theme
287	highly unlikely	722	receive information	1157	useful source
288	little information	723	relatively constant	1158	consider appropriate
289	theoretical model	724	clearly evident	1159	ethical consideration
290	unique opportunity	725	consistent pattern	1160	apply theory
291	technical support	726	positive image	1161	cultural issue
292	urban development	727	potential conflict	1162	demographic characteristic
293	research methodology	728	rich source	1163	economic resource
294	democratic process	729	social institution	1164	key objective
295	seek help	730	systematic study	1165	main element
296	historical development	731	adopt approach	1166	underlying reason

297	modern technology	732	commercial activity	1167	describe method
298	positive influence	733	common ancestor	1168	make recommendation
299	primary responsibility	734	crucial question	1169	provide clue
300	provide access	735	general tendency	1170	deem necessary
301	brief history	736	increasing emphasis	1171	limited resource
302	pivotal role	737	primarily responsible	1172	main characteristic
303	previous section	738	serious challenge	1173	possible consequence
304	complex process	739	theoretical approach	1174	quantitative method
305	empirical study	740	central position	1175	relevant issue
306	critical theory	741	considerable variation	1176	research purpose
307	provide feedback	742	electronic communication	1177	specific characteristic
308	become apparent	743	information flow	1178	varying degree
309	easily accessible	744	natural disaster	1179	follow procedure
310	final product	745	highly influential	1180	make prediction
311	highly skilled	746	research topic	1181	distinct type
312	broad spectrum	747	subsequent analysis	1182	environmental effect
313	diverse group	748	accurate picture	1183	following chapter
314	increasingly complex	749	minimum level	1184	give access
315	international organization	750	overall picture	1185	require consideration
316	primary education	751	previous decade	1186	structural element
317	significant reduction	752	published work	1187	apply method
318	concerted effort	753	similar effect	1188	use format
319	limited access	754	wide variation	1189	contextual factor
320	publicly available	755	central problem	1190	demographic factor
321	significant part	756	crucial factor	1191	general feature
322	effective treatment	757	final outcome	1192	give consideration
323	small fraction	758	original data	1193	multiple source
324	strong relationship	759	general principle	1194	original author
325	advanced technology	760	immediately apparent	1195	practical issue
326	great potential	761	low status	1196	serious consequence
327	obtain information	762	underlying assumption	1197	specific factor
328	brief description	763	considerable research	1198	take initiative
329	extensive research	764	direct involvement	1199	theoretical issue
330	historical evidence	765	natural tendency	1200	conduct interview
331	normal distribution	766	published literature	1201	have consequence
332	critical importance	767	roughly equivalent	1202	publish article
333	dramatic increase	768	alternative means	1203	show trend
334	prior experience	769	interpersonal relationship	1204	use procedure
335	useful tool	770	specific form	1205	deem appropriate
336	available information	771	increased production	1206	social factor
337	ongoing process	772	research effort	1207	specific feature
338	common theme	773	significant variation	1208	additional resource
339	essential component	774	broad definition	1209	economic factor
340	mental state	775	classic study	1210	main principle
341	provide guidance	776	general approach	1211	social consequence
342	recent survey	777	increased productivity	1212	underlying process
343	cultural background	778	specific reference	1213	individual characteristic
344	legal framework	779	historical change	1214	methodological issue
345	little impact	780	make contact	1215	minor change
346	technological innovation	781	particular focus	1216	physical characteristic

347	urban area	782	striking example	1217	technical issue
348	academic research	783	typical example	1218	develop technique
349	powerful tool	784	visual image	1219	available resource
350	significant influence	785	sufficient resources	1220	changing pattern
351	social theory	786	greatly increase	1221	cultural aspect
352	earlier research	787	historical account	1222	individual variable
353	active involvement	788	obvious example	1223	major implication
354	careful consideration	789	alternative method	1224	practical consideration
355	new perspective	790	alternative model	1225	severely affect
356	renewed interest	791	directly responsible	1226	skilled worker
357	technological progress	792	support argument	1227	technical aspect
358	main focus	793	additional cost	1228	theoretical study
359	rural community	794	low profile	1229	contain element
360	significantly affect	795	primary data	1230	develop approach
361	annual rate	796	similar argument	1231	present summary
362	common goal	797	widespread acceptance	1232	show tendency
363	qualitatively different	798	face challenge	1233	basic technique
364	considerable attention	799	basic assumption	1234	face difficulty
365	experimental condition	800	entire period	1235	correct error
366	highly successful	801	find information	1236	encounter difficulty
367	continued existence	802	low percentage	1237	preliminary finding
368	human species	803	previous knowledge	1238	technical detail
369	published research	804	specific problem	1239	encounter problem
370	visual representation	805	reach consensus	1240	make observation
371	causal relationship	806	basic premise	1241	consider relevant
372	essential element	807	essential feature	1242	national boundary
373	overall effect	808	increasing complexity	1243	previous chapter
374	adverse effect	809	original source	1244	adopt procedure
375	broader context	810	public transport	1245	identify factor
376	direct evidence	811	secondary data	1246	cultural factor
377	major challenge	812	social phenomenon	1247	educational qualification
378	highly variable	813	statistical information	1248	financial resource
379	primary objective	814	considerable debate	1249	learning difficulty
380	social contact	815	distinguishing feature	1250	modern method
381	brief overview	816	information retrieval	1251	physical feature
382	central part	817	particularly acute	1252	practical difficulty
383	major change	818	extremely valuable	1253	regional difference
384	quantitative analysis	819	great diversity	1254	social circumstance
385	small proportion	820	increased level	1255	create condition
386	classic example	821	qualitative approach	1256	make judgement
387	critical factor	822	specific purpose	1257	provide illustration
388	global market	823	appropriate action	1258	allocate resource
389	structural change	824	conditional probability	1259	certain characteristic
390	emotional response	825	considerable degree	1260	experimental result
391	evolutionary theory	826	defining feature	1261	preceding chapter
392	industrial development	827	earlier period	1262	relative merit
393	rural area	828	great significance	1263	identify issue
394	small minority	829	particularly appropriate	1264	publish report
395	direct impact	830	related issue	1265	undertake activity
396	environmental damage	831	wider context	1266	achieve goal

397	comprehensive approach	832	complex interaction	1267	provide resource
398	factual information	833	complex issue	1268	structural property
399	assess impact	834	final result	1269	describe procedure
400	negative attitude	835	particularly striking	1270	employ method
401	paramount importance	836	minor role	1271	employ technique
402	highly competitive	837	vital importance	1272	changing circumstance
403	personal interest	838	detailed examination	1273	economic consequence
404	racial group	839	legal obligation	1274	living condition
405	substantial amount	840	thematic analysis	1275	political consideration
406	essential role	841	provide example	1276	urban centre
407	external world	842	economic interest	1277	consider aspect
408	indigenous population	843	statistical test	1278	consider implication
409	major difference	844	widespread belief	1279	have limitation
410	major impact	845	affect development	1280	impose limitation
411	historical data	846	describe process	1281	impose restriction
412	focus attention	847	convey information	1282	require resource
413	alternative approach	848	crucial importance	1283	achieve objective
414	financial management	849	equally valid	1284	appropriate condition
415	negative feedback	850	given period	1285	certain aspect
416	rural population	851	potential benefit	1286	preliminary result
417	verbal communication	852	total income	1287	present difficulty
418	changing nature	853	objective criteria	1288	undergo transformation
419	digital technology	854	set goal	1289	consider impact
420	central theme	855	direct consequence	1290	impose constraint
421	cultural difference	856	homogeneous group	1291	achieve outcome
422	present data	857	integrated system	1292	cover range
423	prime example	858	key concept	1293	create environment
424	western society	859	positive value	1294	identify feature
425	major component	860	precise definition	1295	obtain result
426	technological development	861	reduce stress	1296	quantitative result
427	radical change	862	current technology	1297	similar characteristic
428	sexual contact	863	equally likely	1298	social implication
429	brief discussion	864	large range		
430	explanatory power	865	positive result		
431	special emphasis	866	striking feature		
432	critical review	867	subsequent work		
433	direct relationship	868	underlying structure		
434	fundamental change	869	continuous process		
435	overall performance	870	extremely complex		

Appendix J

Frequency of Collocation Component Words on COCA Academic

Item	Component word 1	Frequency	Component word 2	Frequency
1	significant	60470	difference	63990
2	physical	43749	activity	72591
3	wide	10914	range	27892
4	address	37283	issue	71446
5	religious	29555	belief	10062
6	government	64527	policy	71844
7	social	124430	norm	9795
8	driving	3776	force	23462
9	mutually	2079	exclusive	3019
10	well	90012	documented	4760
11	random	6069	sample	30525
12	ultimate	5263	goal	39016
13	offer	42438	opportunity	35189
14	positively	5118	associated	35506
15	diverse	10763	group	146759
16	high	73147	percentage	11988
17	academic	34419	discipline	11799
18	essential	14669	component	22046
19	fully	12299	understand	72625
20	external	12430	factor	59846
21	close	16205	proximity	2146
22	remain	46939	constant	7425
23	pose	7986	threat	15603
24	prior	17887	experience	52701
25	intrinsic	3268	value	33240
26	widespread	5649	use	126405
27	high	73147	incidence	4846
28	describe	49872	process	86547
29	central	27194	issue	71446
30	ongoing	7383	process	86547
31	research	117851	evidence	36914
32	common	37563	feature	22894
33	appropriate	22560	level	104786
34	experiential	1239	learning	64325
35	newly	5000	formed	7557

36	careful	4459	consideration	12038
37	fundamental	10798	difference	63990
38	major	37314	contribution	14054
39	reach	22962	peak	4241
40	enhance	16400	learning	64325
41	factual	1420	information	78638
42	mutual	4981	understanding	36430
43	initial	16503	phase	15247
44	primary	25002	function	32691
45	high	73147	probability	5456
46	verbal	6494	communication	19416
47	vast	5716	number	70618
48	solar	3732	power	58588
49	keenly	320	aware	9526
50	obtain	24132	result	105209
51	communicate	7843	effectively	9999
52	homogeneous	1404	group	146759
53	seminal	946	work	108156
54	original	16840	research	117851
55	public	79683	welfare	6783
56	roughly	3918	equal	12289
57	free	24839	movement	24473
58	minor	4587	change	89294
59	senior	7212	management	33100
60	dramatic	5518	effect	43884