Development and Initial Validation of a Diagnostic Computer-adaptive Profiler of Vocabulary Knowledge

Mag. phil. Benjamin Kremmel, MA

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

December 2017

To Carol,

because you are who you are,

and I am who I am

Abstract

Vocabulary knowledge is key to the successful use of any language skill (Nation & Webb, 2011) and learning to map a particular meaning to an L2 form for a great number of words is therefore crucial for learners of a foreign language. Vocabulary assessments can play a facilitating role in this learning process, which is why there is now an abundance of assessment tools to measure lexical knowledge. However, few of these tests have undergone sophisticated validation, even after their release into the public domain. Although vocabulary tests are used in numerous pedagogical and research settings, there has been "relatively little progress in the development of new vocabulary tests" (Webb & Sasao, 2013, p. 263). Instead, conventionalized traditions are being reiterated without questioning them. This PhD project has set out to address this gap of an innovative measure of vocabulary knowledge by developing a new diagnostic computer-adaptive measure of form-meaning link knowledge: The Vocabulary Knowledge Profiler.

The present test development project started from scratch by questioning the underlying assumptions and trying to make design decisions based not only on theoretical considerations but empirical evidence. In a series of studies, three major weaknesses of existing vocabulary tests were problematized: (1) selection of item formats, (2) sampling in terms of unit of counting, frequency bands and representativeness, and (3) the general lack of validation evidence and validation models. These issues were explored across four studies in this thesis to design a novel instrument and gather initial validation evidence for it along the way.

The first set of studies presented in this thesis investigated the usefulness and informativeness of different item formats for vocabulary tests and found in a comparison of four different formats that all formats show considerable error in measurement but the MC format may be the most useful because of its systematicity in overestimating scores. The second set of studies found support for the adoption of the lemma as an appropriate counting unit and for a new approach to frequency banding that takes into account the relative importance of frequency bands in terms of the coverage they provide. Based on these foundation studies, test specifications were drawn up and an item bank was created, which was subjected to a large scale trial to admit functioning items to an item pool for creating a computer-adaptive test. A study was conducted to compare two different computer-adaptive algorithms for implementation in the test design, suggesting that a “floor first” design would generate more consistent and representative score profiles. For initial validation evidence, a final study was then conducted to relate scores from the finished test to that of a reading comprehension measure. The findings of the studies presented throughout the thesis are then synthesized to produce an initial version of a validation argument in the structure of Bachman and Palmer’s (2010) Assessment Use Argument to outline both the necessary areas for further research before the launch of the test as well as the collected validation evidence to date that builds a tentative argument that the Vocabulary Knowledge Profiler and of the diagnostic decisions that are made based on its results and use are beneficial to English as a foreign language (EFL) learners and EFL teachers for classroom learning and teaching.

Acknowledgements

I would like to thank everyone who has supported me during my studies and especially during the process of finishing this thesis. I feel incredibly grateful to all of my teachers, mentors and colleagues, from whom I have been privileged to learn.

Firstly, my biggest thanks go to my supervisor Prof Norbert Schmitt for taking me on as his mentee, for his invaluable guidance and support, for innumerable stimulating discussions and for opening more doors for me than I could have ever dreamed of. I hope that I will be able to pay forward his support and provide even half the enthusiasm and mentorship to my future students that he gives to his.

I am grateful to all the members of the Vocabulary Research Group during my time at Nottingham: Hana Almutairi, Samuel Barclay, Dr Mélodie Garnier, Beatriz González Fernández, Duyen Le Thi, Dr Marijana Macis, Dr Ana Pellicer-Sánchez, Niloofar Rahimi, Dr Michael P.H. Rodgers, Dr Kholood Saigh, Dr Pawel Szudarski, Dr Laura Vilkaité and Dr Hilde van Zeeland. Thanks for tolerating my ramblings about language testing, and for the invaluable critical discussions that have made me a better researcher. I would also like to thank the members of the Bilingual Research Group at Nottingham, particularly Dr Kathy Conklin and Dr Walter van Heuven, for everything I was allowed to learn from them about psycholinguistics.

I want to thank Univ.-Prof. Dr. Michael Schratz, Univ.-Prof. Dr. Wolfgang Stadler MA, and in particular Univ.-Prof. Dr. Barbara Hinger MA from the University of Innsbruck for the incredible support and freedom they have given me over the last years. Thanks also to Mag. Sabine Hosp as well as the entire IMoF Team for the great ongoing collaboration.

My very special thanks go to my team of the Language Testing Research Group Innsbruck: Kathrin Eberharter, Sigrid Hauser, Franz Holzknecht, Eva Konrad,

and Matthias Zehentner. Thank you for your friendship and the fantastic teamwork.

Additionally, I want to express my gratitude to Dr Tineke Brunfaut, Dr Luke Harding and Prof J. Charles Alderson from Lancaster University for their ongoing support, encouragement and the opportunities they have opened up for me. I feel extremely fortunate to have you as my mentors and am looking forward to all the collaborative projects we are yet to embark on. I also want to thank Dr Jamie Dunlea, Prof Henrik Gyllstad, Prof Ari Huhta, Prof Barry O'Sullivan, Prof John Read, and Diane Schmitt for providing critical feedback on this project at different stages along the way. Furthermore, I would like to send my very special thanks to all the colleagues, teachers and students that have participated in the studies for this PhD project. This project would not have been viable without your collaboration. I would also like to thank Alistair Shaw for his patience with my programming wishes.

I am thankful to Educational Testing Service and British Council Aptis for their financial support through the TOEFL Small Grant for Doctoral Research in Second and Foreign Language Assessment and the Aptis Assessment Research Award that have both enabled this project to become what it is. The implementation of a computer-adaptive test would not have been possible without their support. In particular, I want to thank Barry O'Sullivan, Jamie Dunlea, Richard Spiby, Vivien Berry, Judith Fairbairn, Mina Patel, and Kevin Rutherford from the Aptis Assessment Research Team for their support and their help with data collection.

My sincere thanks also go to my family for their ongoing encouragement and support and to all of my friends, both those in Austria that made time for me whenever I was around and those in Nottingham that I was fortunate enough to meet and make unforgettable memories with. My special thanks go to Christiane for her unwavering support and for keeping me sane. I couldn't have asked for a more supportive and understanding partner. Thanks for putting up with me.

Finally, I owe thanks, too great to describe, to Carol Spöttl without whom none of this would have been possible. To her, I dedicate this thesis.

Declaration

I declare that the work presented in this thesis is my own and was conducted during my time as a PhD student at the University of Nottingham.

Parts of this thesis have been published in peer-reviewed journals:

- The studies presented in Chapter 3 were published in co-authorship with Norbert Schmitt in *Language Assessment Quarterly* (Kremmel & Schmitt, 2016)

- The studies presented in Chapter 4 were published in *TESOL Quarterly* (Kremmel, 2016)

- The studies presented in Chapters 5-7 are currently prepared for submission in co-authorship with Norbert Schmitt

The chapters in this thesis are more detailed versions of these published papers. The published papers should be used for any citations or page references.

Table of contents

List of figures

List of tables

List of abbreviations

BA	Bachelor of Arts
BNC	British National Corpus
CAT	Computer-adaptive testing
CATSS	Computer-Adaptive Test of Size and Strength
CEFR	Common European Framework of Reference
CITC	Corrected Item-Total Correlation
COCA	Corpus of Contemporary American English
CTT	Classical Test Theory
DIF	Differential Item Functioning
DISCO	Discriminating Collocations
EFL	English as a Foreign Language
FF	"Floor first"
MnSq	MeanSquare
IRT	Item Response Theory
K	1,000
L1	First language
L2	Second language
LFP	Lexical Frequency Profile
LVLT	Listening Vocabulary Levels Test
MC	Multiple Choice
MM	Multiple Matching
MSML	Multi-stage multi-level
NNS	Non-native speaker
NS	Native speaker
PVLT	Productive Vocabulary Levels Test
PVST	Pictorial Vocabulary Size Test
RQ	Research question
SD	Standard deviation
SLA	Second Language Acquisition
SPSS	Statistical Package for the Social Sciences
TED	Test of English Derivatives
VKS	Vocabulary Knowledge Scale
VLT	Vocabulary Levels Test
VLTT	Vocabular Levels Translation Test
VORST	Vocabulary Recognition Speed Test
VST	Vocabulary Size Test
WAF	Word Associates Format
Y/N	Yes/No

1. Introduction

Vocabulary is the "fuel of language, without which nothing meaningful can be understood or communicated" (Gardner, 2013, p. 2). Indeed, learning the vocabulary of a language is therefore probably the key challenge for language learners. In fact, "[m]any learners see second language acquisition as essentially a matter of learning vocabulary" (Read, 2000, p. 1). While this might be an exaggeration, there is certainly merit in the idea that without knowing many words, comprehension and interaction in a foreign language will be difficult, if not impossible. In assisting this learning of vocabulary, vocabulary tests can play a crucial role. They can help identify lexical gaps, facilitate appropriate material selection, and can be useful in monitoring learner's progress to evaluate how well they might be able to meet communicative needs in language-related tasks. Nothing could thereby appear more straightforward than developing and using a vocabulary test. Take some words, ask learners for their meanings, done. Simple enough. Or so it seems.

When investigating the issue of vocabulary assessment more closely, though, a number of questions appear. What is a word? What does knowing a word mean? What is the best way to assess this knowledge? Which words should be selected and how many? How can we interpret the test scores in a meaningful way? Why, or for what purpose, should vocabulary be tested in the first place?

Practitioners and researchers may or may not consider these questions when designing a measure of vocabulary knowledge. They may also choose to simply select and use one of the myriad of vocabulary tests that are publicly available, on- and offline. Too often, however, they then forget to ask themselves how trustworthy, reliable and valid these available vocabulary tests are. The fault is not entirely with these users, though, as test developers of these tests all too often do not provide any information on how they answered these questions themselves. The field of vocabulary assessment seems notorious for a cottage-industry mindset, in which validation evidence is sparse for even the most prominent and most used vocabulary tests, and in which mere assumptions have become unquestioned traditionalized conventions and any "new" vocabulary test seems just another ostinato. This thesis set out to address and challenge some of these assumptions by starting test design from scratch and attempting to base decisions on empirical information wherever possible along the development process. The following chapters will problematize item types, counting units, frequency banding as well as issues of computer-adaptive testing to inform and model state-of-the art validation of vocabulary assessments and suggest possible ways forward in vocabulary testing that should be explored. The thesis exemplifies these issues on the development of a new diagnostic computer-adaptive test of vocabulary knowledge: The Vocabulary Knowledge Profiler.

Chapter 2 will provide a literature review of general key issues in language assessment, such as the concern for test quality criteria, as well as specific issues in vocabulary testing. It will discuss theoretical construct issues of what vocabulary is, and how vocabulary knowledge can be conceptualized to help determine the construct of the new diagnostic measure. The chapter will also provide an overview and critique of existing tests to highlight the need for the new tool to be developed.

Chapter 3 reports on the first foundation study concerned with the informativeness of different item formats in tests of vocabulary breadth. Different frequently used formats were compared against each other in an empirical study to inform the selection of an appropriate response format for the test to be developed. Issues of score interpretation are discussed and an adjustment formula for multiple-choice tests is suggested.

Chapter 4 presents an argument for abandoning the traditional counting unit of word families in favour of the more interpretable unit of the lemma (base form plus inflections). Using corpus analyses, it also argues for a new approach to frequency banding in item sampling and score reporting, which takes into account the relative importance of frequency bands in terms of coverage. The chapter concludes with the proposal of employing narrow bands at the high-frequency end and broader bands at the lower-frequency end for diagnostic usefulness for learners.

Chapter 5 outlines the development of the new measure's test specifications and the diagnostic test items. It also reports on the trialling of the items and the construction of the final item pool for the computer-adaptive test system.

Chapter 6 examines some key issues in computer-adaptive testing and related design decisions for the computer-adaptive implementation of the Vocabulary Knowledge Profiler. In particular, it describes two studies that compared two different adaptive algorithms for their reliability and representativeness.

Chapter 7 provides initial validation evidence in terms of relating the score profiles of the Vocabulary Knowledge Profiler to language skill use. A small-scale study is presented that investigated the vocabulary knowledge profiles of different proficiency groups and probed whether the new vocabulary test managed to distinguish between readers at different Common European Framework of Reference (CEFR) proficiency levels.

Chapter 8 summarizes and synergizes the research presented in the previous chapters into an assessment use argument for the Vocabulary Knowledge Profiler. It discusses claims, warrants and backings for the intended consequences, decisions, interpretations and assessment records of the profiler and points out where additional research was beyond the scope of this PhD project but is needed prior to the launch of the test for a solid validity argument.

2. Literature review

It is widely acknowledged that vocabulary knowledge is integral to success in all language skills (Meara, 1996; Alderson, 2005; Long & Richards, 2007; Daller, Milton, & Treffers-Daller, 2007; Nation & Webb, 2011). Long & Richards (2007) claim that "[v]ocabulary plays an important role in the lives of all language users, since it is one of the major predictors of school performance, and successful learning and use of new vocabulary is also key to membership of many social and professional roles" (p. xii). In particular, scores obtained on various vocabulary tests have been consistently shown to correlate strongly with tests of receptive skills (e.g. Alderson, 2005; Brisbois, 1995; Laufer, 1992; Qian, 2002; Staehr, 2009; Yamashita, 1999)

Although the acquisition of vocabulary has long been viewed a crucial component of language learning and testing, vocabulary research has only gained momentum since the 1990s (Nation, 2011), finally receiving the attention it deserves from applied linguists and language testers. This recognition of the importance of vocabulary knowledge has generated an abundance of assessment tools to measure lexical knowledge. However, few of these tests have undergone sophisticated validation, even after their release into the public domain. Read (2000) therefore rightly cautions us about "making assumptions about what aspect of a language is being assessed just on the basis of the label that a test has been given" (p. 99).

In addition to this dearth of validation research on existing vocabulary tests, despite their being used in numerous pedagogical and research settings, Webb and Sasao (2013) also detect "relatively little progress in the development of new vocabulary tests" (p. 263) and a need for addressing this gap by improving or rethinking ways to assess lexical knowledge. In order to do this, however, the existing literature that has led to the current status of vocabulary assessment must be critically reviewed. This chapter therefore sets out to evaluate the theories and research findings related to the testing of lexical knowledge. It will first briefly discuss key principles in language testing and relate them to the field of vocabulary testing. It will then outline key considerations in vocabulary assessment, particularly pertaining to the construct of vocabulary, conceptualisations thereof and their operationalization in different test formats. The chapter will also analyse the strengths and weaknesses of existing vocabulary tests to identify in detail the gaps that this PhD thesis aims to address.

2.1.Key issues in language testing

Since any test of lexical knowledge is essentially a language test, the core quality principles of language testing also apply to this very specific type of measurement instrument. The following section will outline these principles and will evaluate to what extent each applies to the measurement of lexical knowledge.

Bachman and Palmer (1996) state that a language test's usefulness is a function of six quality criteria: construct validity, reliability, authenticity, interactiveness, impact and practicality. Of these, however, reliability and validity are regarded as the "essential measurement qualities" (Bachman & Palmer, 1996, p. 19). In a more recent model for validation, the Assessment Use Argument, Bachman and Palmer (2010) introduce a number of new criteria, which they argue to pertain to the claims and warrants in the use argument structure of an assessment tool or system. These terms, although deliberately trying to avoid the previously suggested and somewhat loaded terminology, do overlap significantly with most of the criteria generally established for judging a test's usefulness. Their criteria of beneficence, value sensitiveness, equitability, meaningfulness, impartiality, generalizability, relevance, sufficiency and consistency are, in essence, very similar to the concepts of traditional models such as impact, construct validity, content validity or reliability. It remains debatable whether their terminology really adds to the validation discussion, particularly since their ultimate criterion of beneficence seems very problematic (Fulcher, 2015). Also, since they suggest these criteria within their framework of communicative language tests, it appears questionable whether all of these criteria apply to the measurement of lexical knowledge in equal fashion. While validity, reliability and practicality are certainly also key to vocabulary assessments, the role of authenticity, interactiveness and impact might be slightly different in vocabulary tests than in skill tests.

Reliability is typically defined as “consistency of measurement” (Bachman & Palmer, 1996, p. 19) and is a crucial characteristic of any useful test. While it will be argued in more detail later in this thesis why an overreliance on the Cronbach alpha value, the traditional indicator of internal test consistency and often just referred to as the value indicating the “reliability” of a test, may be problematic particularly for vocabulary measurements, it is undisputed that the concept of measurement consistency is pivotal for all tests.

Authenticity, defined as the degree of correspondence between characteristics of TLU [target language use] tasks and test tasks (McNamara, 2000), thus might not be of prime concern in vocabulary tests. As “a means for investigating the extent to which score interpretations generalize beyond performance on the test to language use in the TLU domain“ (Bachman & Palmer, 1996, p. 24), the principle seems to be more important for tests of language skills than this specific area of linguistic knowledge. However, this only holds with the assumption that there is only one kind of vocabulary test. The principle does become important to differing degrees depending on the type of vocabulary test and potentially also the context, in which lexical items could be presented in a test.

Similarly, interactiveness seems to pertain to skills tests more than vocabulary tests at first glance, and it possibly does for the most part. However, vocabulary tests similarly need to account for “extent and type of involvement of the test taker‘s individual characteristics in accomplishing a test task“ (Bachman & Palmer, 1996, p. 25) or item, i.e. the effect of factors such as age, gender, motivation and L1.

A testing principle key to any language test and indeed also to vocabulary testing is that of practicality. Described as the balance between available and required resources (Bachman & Palmer, 2010), this issue is of prime importance as one would want a measure as detailed and reliable as possible but has to bear time and financial constraints in mind. Vocabulary test designers therefore often have to consider the trade-off between the number of lexical items they wish to target and the amount of knowledge information they strive to attain for each of those targets.

The prime concern for tests of lexical knowledge, however, must be, as for any language test, the overarching notion of validity. Bachman (1990) states that validity is "the most important quality of test interpretation or use" (p. 25) and that therefore validation is "the primary concern in test development and use" (p. 236). Alderson, Clapham and Wall (1995) echo this by claiming that validity is "the most important question of all in language testing" (p. 170). Most validation researchers in language testing hold that validity is not a property of an assessment instrument itself, but describes "an integrated judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). Messick (1989) stresses that "key issues of test validity are the interpretability, relevance and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use" (p. 13).

In its simplest conceptualization, validity refers to whether a test "measures accurately what it is intended to measure" (Hughes, 2003, p. 26). This realist view sees test validity as a psychometric property of a test itself rather than of an interpretation (Borsboom, Mellenbergh, & van Heerden, 2004), which, in its extreme form, has also been pointed out to be problematic (Fulcher, 2014). However, the notion of validity has been defined considerably differently by different scholars. It is thus important for any assessment instrument to outline which idea of validity it employs as this is the basis for both the claims and the resulting needs for validation evidence.

Cronbach and Meehl (1955) postulated four types of validity as separate entities: "predictive validity, concurrent validity, content validity, and construct validity" (p. 281). The latter two have thereby been profoundly influential in measurement theory. Their definition of content validity as the extent to which a test samples adequate and representative measures "of a universe in which the investigator is interested" (Cronbach & Meehl, 1955, p. 282) still features strongly in many contemporary views on validity (McNamara, 2000). Their introduction of the term "construct validity", however, was crucial as it nowadays lies at the heart of many conceptualizations of validity.

Cronbach and Meehl (1955) describe construct as "some postulated attribute of people, assumed to be reflected in test performance" (p. 283). Construct validation is thus "involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined'" (Cronbach & Meehl, 1955, p. 282). Lado's (1961) concern thereby still holds that for a test's score interpretation to be accepted as valid, the test must measure very little or nothing else than that particular attribute or quality it purports to measure. For vocabulary tests, this implies a need to establish some distinction between the testing of lexical knowledge and the testing of other language skills or knowledge areas, which is arduous and seemingly impossible (Read, 2000). It could, for instance, mean that the involvement of other language skills like reading or writing should be kept to a minimum if one is truly only interested in a person's vocabulary knowledge. The crux with this, however, is that it depends very much on the conceptualization of what vocabulary knowledge, and thus the construct, is. It will be demonstrated in Section 2.2.1 that this is far from agreed upon.

Construct validity is also the key consideration at the core of Messick's (1989) unified validity concept. His seminal framework is still one of the most prominent notions of validity or is, at the very least, crucial to understanding all current notions of validity. The framework "highlights the important, though subsidiary, role of specific content- and criterion-related evidence in support of construct validity in testing applications" (Messick, 1989, p. 20). Messick's reconceptualization of validity as a multifaceted amalgamate also resulted in a paradigm shift in terms of validation procedures. He maintained that the different aspects of validity called for an expansion of methods of gathering evidence for establishing various aspects of validity. In other words, only a combination of different categories of validity evidence adequately reflects the value of a test for a stipulated purpose (Messick, 1989). This is particularly relevant for vocabulary testing as the traditionally heavily psychometrically-based view of validity (Cronbach & Meehl, 1955; Lado, 1961) means that many vocabulary researchers still seek to validate their tests through correlation studies. Messick, however, claims that "different inferences from test scores require different blends of evidence" (Messick, 1989, p. 49), to eventually contribute to establishing construct validity. Unfortunately, only a few studies have been conducted that offer such a blend of evidence for existing vocabulary tests.

Crucial to any test's validation is thereby the purpose of the test. Henning (1987) rightly maintains that "the term valid when used to describe a test should usually be accompanied by the preposition 'for'. Any test then may be valid for some purposes, but not for others" (Henning, 1987, p. 89). Bachman (1990) claims that "to refer to a test or test score as valid, without reference to the specific ability or abilities the test is designed to measure and the uses for which the test is intended is therefore more than a terminological inaccuracy" (p. 238). However, past and current practice in vocabulary test design and use appears to frequently neglect this factor, jeopardizing the (construct) validity of findings and claims.

Recent models of validity seem to devalue the role of construct validity due to the complexity involved in describing linguistic constructs. Also, the lack of concrete practical guidance as to how to gather construct validity evidence in Messick's approach has been criticised by language testing researchers (Kane, 2012). New theories of validity have therefore put the validation procedure at their centre, downplaying the need for a definition of the theoretical construct (Chapelle, 2012).

Kane's validation argument is now seen as an "alternative standard framework for thinking about validity in language testing" (McNamara, 2006, p. 47). Kane's argument-based approach to validity attempts to overcome problematic aspects of methodologically operationalizing theories of validity (Kane, 1992, 2004). The main focus thus lies on systematically identifying threats to validity a priori and developing procedures that support proposed score interpretations (development stage) and investigating them in later validation studies to critically evaluate the plausibility and appropriacy of the proposed interpretations and uses of scores (appraisal stage) (Kane, 2012). Kane's argument-based validation employs an interpretive argument, specifying postulated score uses and interpretations by outlining "a network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the assessment scores" (Kane, 2012, p. 8), and a validity argument that evaluates the coherence, clarity and plausibility of that interpretive argument. The broader the network of assumptions, inferences and generalizations, the more validation evidence is required to assert the legitimacy of that interpretation. Based on Toulmin's (1958) argument framework, Kane's interpretive model consists of inferences from given data to form claims. These claims need to be justified by warrants, which are again substantiated to assert that the warrants and inferences are legitimate and appropriate (Kane, 2012).

The interpretive argument thus seems to replace the construct definition as basis for validation. However, at a closer look, one can see that even this seemingly pragmatic model does not get by without reference to some predefined construct, be it ever so vague. Figure 1 below, outlining a model interpretive argument and its elements and interrelations, appears to indicate this clearly.

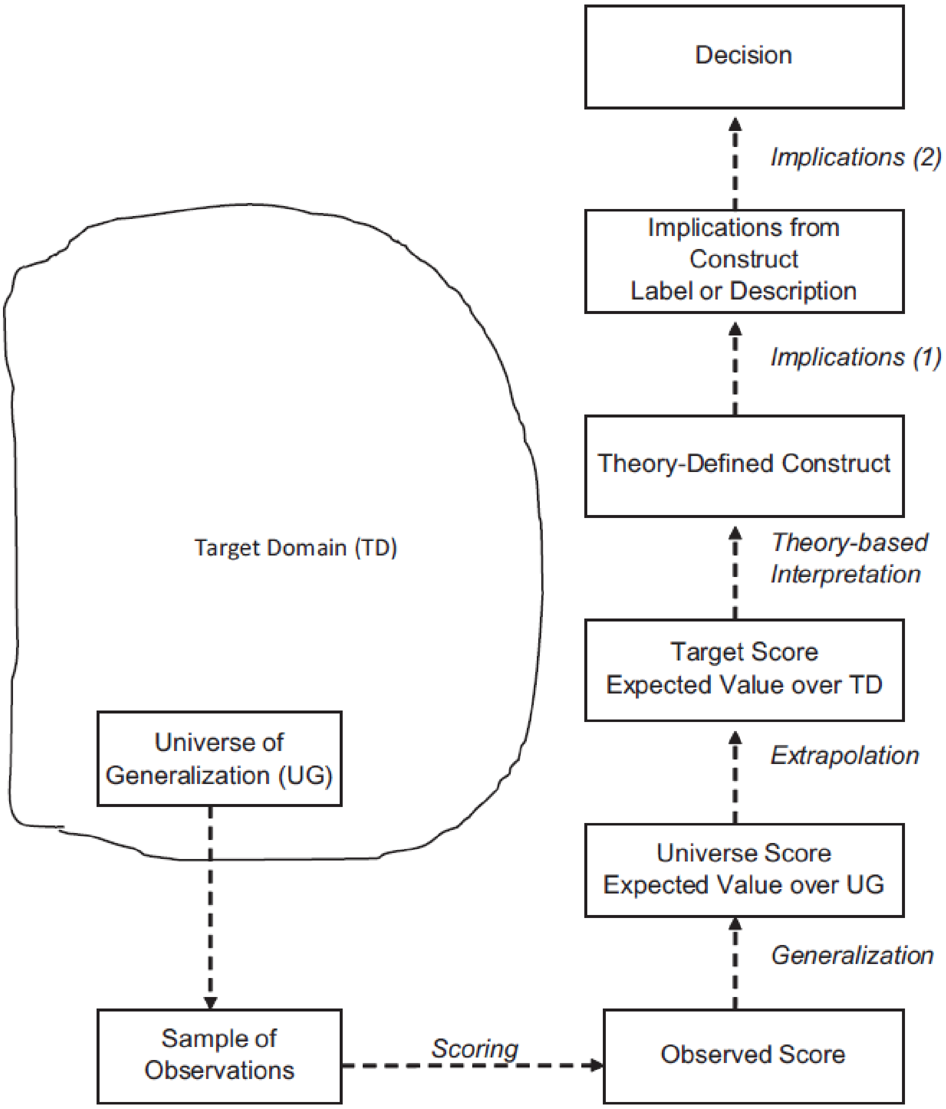

Decision
Implications (2)
Implications from Construct Label or Description
Implications (1)
Theory-Defined Construct
Target Domain (TD)
Theory-based Interpretation
Target Score Expected Value over TD
Universe of Generalization (UG)
Extrapolation
Universe Score Expected Value over UG
Generalization
Sample of Observations
Scoring
Observed Score

Figure 1: Elements and context of an interpretive argument (Chapelle, 2012, p. 21)

Even the first inference "scoring" advances the argument "with a warrant that the observations of performances on a test are scored accurately, appropriately with respect to the *construct* measured [...]"(Chapelle, 2012, p. 20, emphasis added). Generalization, extrapolation and particularly theory-based interpretation and implication certainly seem implausible without making recourse to a construct. As useful as Kane's model thus seems in practical terms, it is still grounded in principle on the validity aspects put forward by Messick, which implies that there is still a need for a construct defined as clearly as possible. Fulcher (2015) offers a useful and convincing critique of such argument-based, instrumentalist validation models. He suggests to overcome these relativistic utilitarian validity models (as well as innately relativistic, postmodern constructionist approaches) by moving towards a more Pragmatic (with a capital P) realist view of validation. Fulcher (2015) also exposes technicalist models such as Weir's (2005) as mere checklist approaches that generate "'marginally relevant' information from a 'do-it-yourself' kit of disjoint facts" (p. 119) and suffer from the "'gilding the lily' fallacy" (p.119). He further argues for a Pragmatic Realism over a purely realist view such as the one taken by Borsboom et al. (2004), which regards validity as a property of a test rather than an interpretation. In Fulcher's view, the notion excludes contingency completely and assumes "a viewpoint of providence" (2015, p. 123) that makes it "just a touch too arrogant" (p. 123) because it implies "some immediate a-historical insight into the nature of reality" (p. 122).

Instead, Fulcher (2015) proposes a Pragmatic Realism in validation that is based in experience (linguistic data or observation of communication), is optimistic, and which acknowledges a degree of contingency. It combines data-driven and effect-driven aspects in that it references a test to the criterion of language use (in terms of test content, scoring and inferences) and takes the social nature of testing into account by explicating the test purpose clearly, not just as an addendum in the validation process, but articulating it at the very start of test design and development. Fulcher pointedly asks "How can we develop a good test if its purpose isn't clearly articulated?" (Fulcher, 2015, p. 126). For this purpose, he defines a construct as "the abstract name for a complex idea derived from observations of co-occurring phenomena, the purpose of which is to explain the coherence of our perceptions and make predictions about the likelihood of future states or events" (Fulcher, 2015, p. 130). Although his suggestion of criterion-referenced validation based on careful and extensive domain analysis appears reasonable for communicative language tests, Fulcher (2015) does not provide detailed description or guidance as to how to operationalise a Pragmatic Realist validation approach, particularly for diagnostic vocabulary tests. It could, however, be argued that the domain analysis could take the form of sampling target items from a relevant and well-balanced corpus. Also, criterion-referencing could be achieved by comparing vocabulary test results to language use in the actual language skill that the test claims to be related to. For instance, a test of written receptive vocabulary could provide the scores with meaning by looking at how it relates to candidates' ability to employ the word knowledge in actual written reception, i.e. reading. This, however, makes it still indispensable to discuss and outline the construct of vocabulary tests in terms of what a word is and what it means to know a word. This will be addressed in the following section.

2.2.Key issues in vocabulary testing

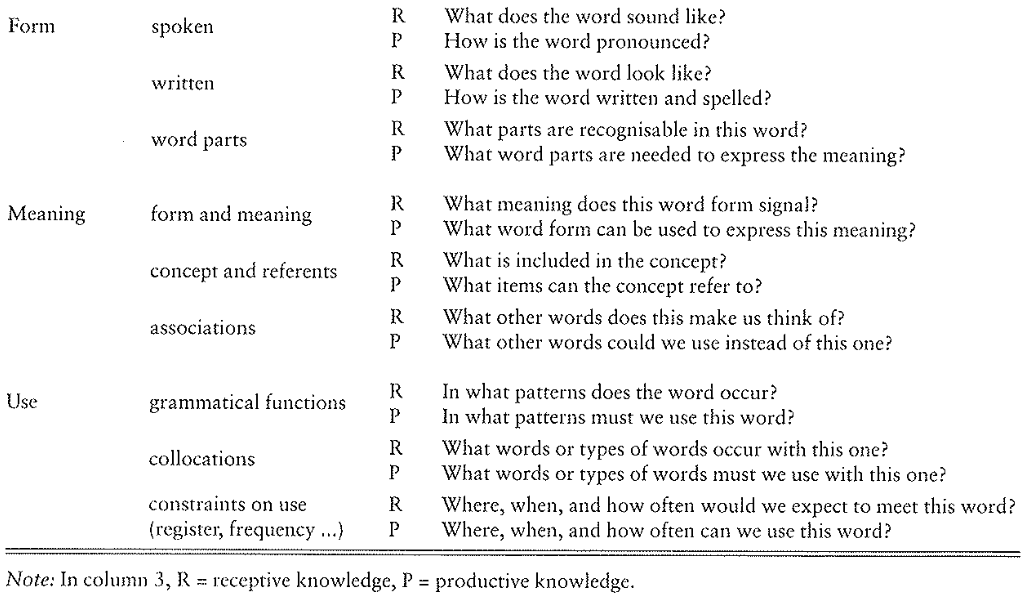
2.2.1. Construct definition – What is vocabulary knowledge?

According to most validation theories, the construct of a test needs to be determined before any test design or indeed validation can take place. In the context of measuring vocabulary knowledge, it is thus essential to define both "vocabulary" and "knowledge" thereof as clearly as possible. Laufer and Goldstein (2004) stipulate that "[v]ocabulary tests are contingent upon the test designer's definition of lexical knowledge" (p. 399). However, Read and Chapelle (2001) maintain that the nature of vocabulary as an assessment construct is "ill-defined" (p. 1) as different scholars have chosen and continue to choose different perspectives and approaches to the issue at hand. While a certain variety of approaches is in itself not highly problematic, though undesirable for comparability of studies, it poses considerable challenges when researchers' assumptions about the nature and scope of lexical knowledge are only implicitly alluded to or not clearly outlined at all. It seems, however, that vocabulary researchers have so far "given comparatively little attention to defining 'vocabulary knowledge' or 'vocabulary size' as theoretical constructs" (Read & Chapelle, 2001, p. 7) that form the basis of test selection and construction.

In everyday conversation, there is a tendency to think of vocabulary knowledge "as an inventory of individual words, with their associated meanings" (Read, 2000, p. 16). Hill (2000) also observes that "vocabulary" is all too often equated with individual words. Put differently, "if you 'have a big vocabulary' you 'know a lot of words'" (Lewis, 1993, p. 89). As such, one would think it should not be too complicated to measure vocabulary knowledge (Read, 2007). However, the seeming simplicity soon falls apart at a second look (Miller, 1999).

Zhang and Anual (2008) claim that "it is difficult to reach a consensus on what is involved in word knowledge and how to measure vocabulary knowledge due to the complexity of the construct of what it means to know a word" (p. 55). There is no agreement among applied linguists as to what constitutes a word (Read, 2000). Numerous researchers tend to define vocabulary as "words", or at least use the terms synonymously (Lewis, 1993; Thornbury, 2002), and Moon (1997) agrees that "it is natural to focus on the word as the primary unit" (p. 40) when looking at vocabulary. However, researchers have questioned for more than two decades now whether it is "sufficient to equate 'vocabulary' with single words" (Schmitt & McCarthy, 1997, p. 1). The notion that language is also made up of formulaic multi-word chunks that are stored similarly to individual words is backed by findings from computerized corpora (Sinclair, 1991) but is still often ignored in the measurement of vocabulary knowledge. Sinclair (2004) claims that "so strong are the co-occurrence tendencies of words, word classes, meanings and attitudes that we must widen our horizons and expect the units of meaning to be much more extensive and varied than is seen in a single word" (p. 39). This suggests that traditional tests of vocabulary paint only half the picture as they do not take into account sequences, "continuous or discontinuous, of words or other elements, which [are], or appear[s] 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" (Wray, 2002, p. 9). For the integration of formulaic sequences in the construction of vocabulary tests, however, two key issues pose considerable problems. The first is conceptual in nature in that there is currently no agreed upon definition or classification of formulaic sequences. This renders it almost impossible to create frequency-based lists of vocabulary items, incorporating both single words as well as multiword units of different kinds, from which a test developer could sample items. The second issue is one of practicality. Creating such a list, even with selected agreed upon categories of formulaic sequences, would be beyond the remit of many test development projects like the present one. And if existing frequency lists were simply combined, it would still be unclear at what rate single words and formulaic sequences should be sampled per frequency band in order to arrive at a representative sample of both. For these reasons, the present test development project will also bracket out the problem of formulaic sequences despite the awareness of the limitations this implies for the final product. In light of these and other issues that will be explored in Section 2.2.3, it appeared more important to problematize some of the more basic issues before moving on to complex conundrums, such as the incorporation of formulaic sequences.

In addition to ambiguity in the field regarding the definition of a word and a resulting ambiguity as to which form of lexical unit to include in vocabulary tests, researchers also differ in their conceptualizations of what is involved in knowing a word. It seems to be agreed upon that knowledge frameworks are useful for both vocabulary teaching and testing (Schmitt, 1995). Following Richards' (1976) early description of vocabulary knowledge, Nation (2001) proposed what is perhaps currently the most influential and comprehensive framework of aspects of word knowledge.



Form	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognisable in this word?
		P	What word parts are needed to express the meaning?
Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	constraints on use	R	Where, when, and how often would we expect to meet this word?
	(register, frequency ...)	P	Where, when, and how often can we use this word?

Note: In column 3, R = receptive knowledge, P = productive knowledge.

Figure 2: Nation's (2001) framework of word knowledge

Nation's (2001) multidimensional view of vocabulary knowledge is divided into the three categories form, meaning and use. These are further detailed in the framework in that each of these components and subcomponents of word knowledge need to be "known" at both the receptive and the productive level in order to achieve full mastery of a lexical item.

Nation tries to capture and simplify into a manageable taxonomy what is not as clear cut as it may seem. This is illustrated by the overlap of the two macro-components form and meaning in the first subcategory of meaning. Also, it remains unresolved how, for instance, collocations that function as phrasemes with their own distinct meaning as potentially polysemous single-meaning units, can be placed within this frame.

In other words, the categories, while seemingly theoretically sound, pose problems in real-world application as they are, albeit to varying degrees, interrelated and mutually interdependent. Some components such as "concepts and referents" seem more difficult to grasp in practical terms for test construction and less clearly defined than others. Nation's scheme also suffers from the weakness that it does not specify whether it is an implicational scale that is presented here. While it may well be that the "use" of a word constitutes a higher form of word knowledge than, for example, spelling, this may not hold true of other relations between subdimensions. The relation between receptive and productive word knowledge has also been problematized by other scholars (Melka, 1997). Most of the weaknesses of this framework, thus appear to be due to the lack of a comprehensive theory of vocabulary development (Schmitt, 2010). One could further argue that some of the subcomponents could even be elaborated further as words might have different sounds in different (regional) contexts, etc. However, despite its shortcomings, this framework seems to be the most thorough and useful view of lexical knowledge to date.

Based on multidimensional views of vocabulary knowledge, many vocabulary assessment researchers have attempted a clearer definition of vocabulary knowledge for assessment purposes by distinguishing between two dimensions of vocabulary knowledge: depth and breadth. While breadth denotes to the quantitative size of a person's knowledge of lexical items (Lewis, 1993), depth refers to "how well one knows a word" (Qian, 2002, p. 515), subsuming "such components as pronunciation, spelling, meaning, register, frequency, and morphological, syntactic, and collocational properties" (ibid.). In a first elaborate definition, Anderson and Freebody (1981) distinguish between "breadth, by which we mean the number of words for which the person knows at least some of the significant aspects of meaning" (p. 92) and quality or depth, referring to "all of the distinctions that would be understood by an ordinary adult under normal circumstances" (p. 93). Despite its elaborateness, however, the definition remains vague and challenging to operationalize in vocabulary tests, particularly tests of depth. Schmitt (2014) recently also makes the point that the "diversity of depth conceptualizations makes it extremely difficult to know how to approach depth from a theoretical perspective" (p. 915). He concludes that "there can be no clear distinction between size and depth" (p. 942) as all aspects of word knowledge are to some, yet undetermined, extent interrelated and even testing only the form-meaning link in a size test is already a measure of, arguably very shallow, depth.

While breadth tests have been criticized for only providing a superficial indication of how well words are known (Read, 2000), depth or quality of knowledge tests focus on more than merely the most common meaning of a target word or a single synonym (Dolch & Leeds, 1953). Depending on their operationalization of depth they allow for testing of additional, even figurative meanings and also finer-grained partial knowledge, similar to what can be probed for in more laborious interview tests. This is crucial as Schmitt (2008) argues that depth of knowledge is essential to understand and use a word appropriately.

Breadth and depth, though used in a dichotomous fashion, are not completely unrelated or independent of each other. Research by Nurweni and Read (1999) found that there is a relationship between breadth and depth, although it seems to depend on a learner's proficiency level. In their study, the correlation between the lexical breadth measure and the lexical depth measure was stronger for high proficiency learners than for low-proficiency learners. Qian (1999) also reports positive correlations between scores on a breadth measure (Nation's Vocabulary Levels Test) and a depth measure (Read's Word Associates Test). A strong relationship between learners' breadth and depth measure scores, even for lower proficiency candidates, was also found by Vermeer (2001). However, Greidanus et al. (Greidanus, Bogaards, van der Linden, Nienhuis, & de Wolf, 2004) doubt that we can go so far as to say that "there seems to be no conceptual distinction between breadth and depth" (Vermeer, 2001, p. 222). Rather, Nation and Webb (2011) agree with Qian (1999) that the correlation might be due to a partial overlap in measures as both often contain a semantic or form-meaning component.

While vocabulary breadth seems fairly straightforward to conceptualize, often referring to the number of words for which the form-meaning link is known, depth of vocabulary knowledge appears more complex to define. This shows in the various approaches taken to explore it. Read (2004) therefore states that the single term "depth" might be misleading. He suggests to use the three more specific terms (1) precision of meaning, (2) comprehensive word knowledge and (3) network knowledge instead. Precision of meaning thereby refers to an elaborate, specific knowledge of a word's meaning(s) that goes beyond a merely vague idea of what it means. Comprehensive word knowledge delineates knowledge of a words' semantic, "orthographic, phonological, morphological, syntactic, collocational and pragmatic characteristics" (Read, 2004, p. 211). Word knowledge aspects of Nation's (2001) seminal framework would fall under this category of depth. Thirdly, network knowledge pertains to the integration of a word into the mental lexicon and "the ability to link it to - and distinguish it from - related words" (Read, 2004, p.212). Schmitt (2014) identifies up to seven categories of conceptualizations and operationalizations of depth that might aid in describing this intricate construct more clearly than when only speaking about depth in vague terms.

Within the depth or quality of knowledge approach, Schmitt (2010) observes another important classification. Following Read (2000), he distinguishes developmental approaches, "describing the incremental acquisition of a word along a continuum of mastery" (Schmitt, 2010, p. 216), from dimension or components approaches that specify different kinds or types of word knowledge (Schmitt, 2010). While developmental approaches, often in the form of scales, account for the undoubtedly incremental nature of vocabulary learning, their operationalization seems currently almost impossible given the little knowledge we have about how vocabulary develops exactly. Schmitt (2010) states that "[v]ocabulary acquisition theory is not advanced enough to guide the creation of a principled developmental scale" (p. 217) at this point. Even though such vague scales might bear some merit for language pedagogy, they seem of limited usefulness in terms of vocabulary assessment, as it is yet to be demonstrated where such a scale should begin or end and, indeed, how many and which stages would lie in between these two points (Schmitt, 2010). Schmitt's (2010) speculation that there might be an uncountable number of small knowledge increments render it questionable whether reasonable and generalizable developmental stages can be identified at all.

Many researchers have thus attempted to operationalize a componential approach to measuring quality of lexical knowledge. According to Schmitt (1998), the advantage of such an approach is that it could provide a comprehensive and rich, though time-consuming, measurement of vocabulary knowledge. Also, breaking vocabulary knowledge down into separate dimensions or components might make their assessment more manageable and diagnostically more valuable for score users. If further allows for investigations and hypotheses about interrelations of separate components, which might, at best, even be hierarchical to some extent. This, however, is yet to be demonstrated. The potential comprehensiveness of dimension approaches are at the same time both their biggest appeal and drawback. With vocabulary knowledge being as multifaceted as established, it seems impossible to measure all aspects of this knowledge in one test. Also some components, for instance register (Schmitt, 2010), might be very difficult to test at all. Read (2000) states that even if several dimensions were to be tested, "there is a danger of finding out more and more about the test takers' knowledge of fewer and fewer words" (p. 248), which might only be useful for a very limited number of purposes.

This might be one reason why most instruments that measure vocabulary knowledge only focus on one dimension or component (Qian & Schedl, 2004), neglecting a comprehensive view of all vocabulary knowledge, mostly for practical reasons. Despite the acknowledgement and influence of multifaceted views of lexical knowledge in applied linguistics research, most available vocabulary knowledge tests still focus predominantly on solely one facet, the quantity of learners' vocabulary knowledge. This often results in measurements of the form-meaning link only. While this has been established to be the most crucial of all aspects of word knowledge for language learners (Laufer, Elder, Hill, & Congdon, 2004; Schmitt, 2008), it still seems questionable whether tests measuring only this dimension provide a sufficient representation of lexical knowledge for meaningful score interpretation. A balanced measure that accounts for breadth and some depth would thus seem an important contribution to the field.

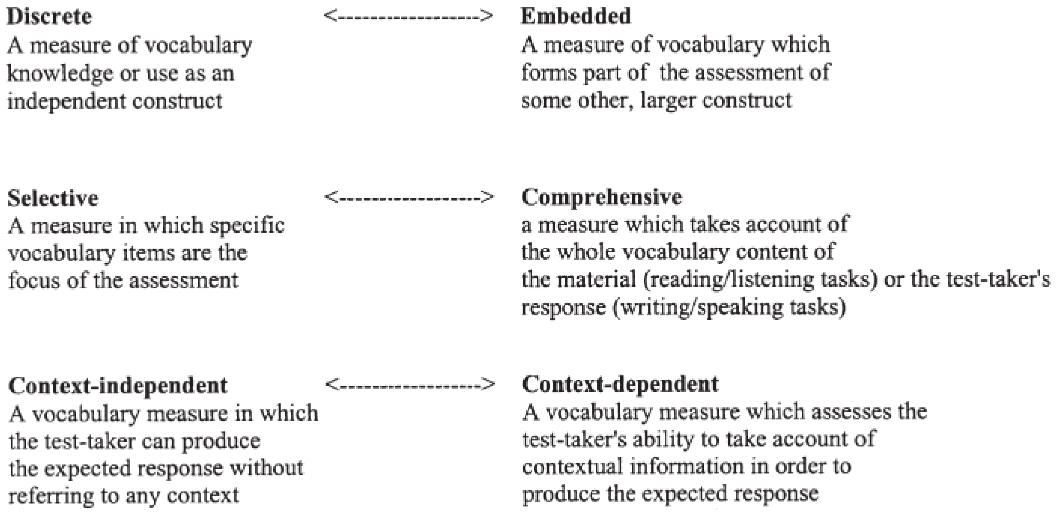
The matter, however, is further complicated when taking into account that some vocabulary researchers have proposed three dimensions of vocabulary knowledge. Such an approach can, for instance, be found in Daller, Milton and

Treffers-Daller's (2007) concept of "lexical space", which comprises breadth, depth and fluency or automaticity of retrieval. Laufer and Nation (2001), as well as Zhang and Lu (2013), have argued for including supplementary fluency measures in vocabulary knowledge measures for a more complete picture of learners' lexical abilities. However, lexical decision tasks seem to suffer from the same weaknesses as checklist tests (Pellicer-Sánchez & Schmitt, 2012), and minimal validation has been carried out on other computerized tests that measure speed of retrieval, such as the VLT-based Vocabulary Recognition Speed Test (VORST) (Laufer & Nation, 2001). According to Laufer and Goldstein (2004), "strength" of vocabulary knowledge, distinguishable from breadth and depth, is a further dimension to be considered in framing and measuring word knowledge. Other scholars even proposed four dimensions of lexical knowledge: size, depth, connection or organization, and speed of lexical access (e.g. Read, 2004b; Schmitt, 2010), rendering the idea of a single comprehensive measurement instrument for vocabulary knowledge almost impossible.

A further important distinction is suggested by Henriksen (1999). Her tripartite model of vocabulary dimensions is again specified in several subcomponents. The first dimension focuses on the continuum of partial-precise knowledge, onto which vocabulary items of different tests can be placed. The second dimension, conceptualized as a network rather than a single cline, refers to depth of knowledge and subsumes different types of knowledge as outlined, for instance, in Nation's (2001) aspects of word knowledge. Dimension three relies on the distinction between receptive and productive knowledge, which has, however, also been challenged (Melka, 1997) and might indeed be a very intricate matter of different aspects of word knowledge for each individual lexical item being known to various receptive and productive degrees (Schmitt, 2010).

2.2.2. Vocabulary assessment frameworks

As hinted at in the previous section, different approaches have been taken to operationalizing the construct vocabulary knowledge, or parts of it, in vocabulary tests. Read (2000) outlines three dimensions of vocabulary assessment which he stipulates as continua. According to this typology, vocabulary measures can be classified in terms of their degree of discreteness, selectiveness and context-dependency.



Discrete A measure of vocabulary knowledge or use as an independent construct	<----------------->	**Embedded** A measure of vocabulary which forms part of the assessment of some other, larger construct
Selective A measure in which specific vocabulary items are the focus of the assessment	<----------------->	**Comprehensive** a measure which takes account of the whole vocabulary content of the material (reading/listening tasks) or the test-taker's response (writing/speaking tasks)
Context-independent A vocabulary measure in which the test-taker can produce the expected response without referring to any context	<----------------->	**Context-dependent** A vocabulary measure which assesses the test-taker's ability to take account of contextual information in order to produce the expected response

Figure 3: Three dimensions of vocabulary assessment (Read, 2000, p.9)

In the first dimension, he distinguishes between discrete and embedded measures of vocabulary at the extreme ends of the cline. Discrete measures thereby postulate lexical knowledge as an independent, distinct construct, traditionally viewing lexical knowledge as some sort of latent trait (Read & Chapelle, 2001). While Read (2000) maintains that most vocabulary tests to date employ this assumption, this might not necessarily be the case when taking a closer look. It certainly seems that this claim is valid prima facie but probably only because embedded measures are, because of their integrated nature, rarely identified as vocabulary tests of their own right. Since embedded measures usually assess vocabulary as part of a larger construct, say for instance as one rating criterion in a scale and setting where writing or speaking ability is tested, they are somewhat covert vocabulary measures which are often neglected in the academic discourse on vocabulary tests. It seems worth noting that, according to Read's (2000) taxonomy, the discreteness or embeddedness of a vocabulary measure is not related to the form of presentation of lexical items in a test. A test may be discrete regardless of whether or not the target items are presented in isolation. Even a test that presents words in a substantial amount of context may not be considered embedded if the questions are targeted at individual lexical units and the scores on the test are not interpreted as indicators of, in this case, reading ability. Embedded measures have the advantage of being more authentic and integrating an element of vocabulary use. However, they appear difficult to score reliably as they potentially muddy the measurement (Weir, 1990) due to the many other factors that play into such types of assessment, but which are challenging to control for (Schmitt, 2010).

The second cline suggested by Read (2000) spans from tests being selective in their character, i.e. focusing on specific lexical items that are tested according to principled preselection, to comprehensive assessment instruments, which take "account of the whole vocabulary content of the input material (reading/listening tasks) or the test-taker's response (writing/speaking tasks)" (Read, 2000). Most vocabulary tests are selective in nature (Read, 2000), which might be due to them being traditionally based on a trait view of vocabulary knowledge and thus often being discrete measures. It might, however, also be due to the fact that the target words can be carefully selected rather than be subject to holistic judgments. Selective measures give the test developer control as the sampling can be based on a principled rationale, for instance by using frequency as selection criterion. As will be discussed later, not all purportedly selective measures are equally successful in exerting this degree of control. However, it does seem that the amount of control that selective measures allow for is also an advantage in terms of the comparability of scores.

Dimension three relates to the context in which target words are presented. On a vocabulary test, lexical items might appear in isolation or within the context of one or several sentences. However, a test's position on this dimension is not defined by the mere presence of context but rather by the question "to what extent the test takers are being assessed on the basis of their ability to engage with the context" (Read, 2000, p. 11) when answering items. If candidates need to make use of contextual information to arrive at the correct answer, a test can be considered relatively context-dependent.

Discrete tests can thus be either context-dependent or context-independent, while embedded tests tend towards the context-dependent end of the scale as they usually assess a candidate's ability to use vocabulary appropriately in a particular cotext and context. Schmitt (2010) suggests that context-dependent formats might be more useful for tapping into contextualized aspects of word knowledge, such as collocation.

Although Read (2000) only links the first of these three dimensions to the construct of vocabulary tests, it could be argued that, particularly in a contemporary understanding of validity, all three pertain to the construct that is to be measured. A test's position on Dimension 3 most certainly has an impact on the test's construct or rather vice versa. The second dimension could be seen as relating to content validity, which forms a core component of construct validity in a Messickian view.

Read's (2000) tripartite model was further developed into a broader classification framework by Read and Chapelle (2001). They distinguish three types of construct definitions in vocabulary assessment: (1) trait definitions, (2) behaviourist definitions, and (3) interactionalist definitions. All of these operationalize the abovementioned components to varying degrees along the three outlined continua. Researchers subscribing to trait definitions in vocabulary testing are primarily concerned with vocabulary knowledge "as a trait without reference to any particular context of use" (Read & Chapelle, 2001, p. 8). Test performance is thus solely attributed to the knowledge characteristics of the individual learner, resulting conventionally in the presentation of vocabulary test items in a discrete, selected, isolated and context-independent fashion. Behaviourist definitions, by contrast, stand in line with so-called performance testing traditions. According to Read and Chapelle (2001), they rely on specifying the context in which language is used as they assume that vocabulary, or any other aspect of linguistic knowledge for that matter, cannot be singled out for discrete scoring as this underlying knowledge is too elusive to warrant precise definition. Synthesizing these two extreme positions, interactionalist approaches hypothesize a "context-specific underlying ability" (Read & Chapelle, 2001, p. 9), i.e. a trait manifested in a particular usage context. In tests following such an approach, vocabulary is often tested in an embedded, comprehensive and context-dependent manner with, for instance, the rating scale of a written performance specifying vocabulary range or accuracy as a marking criterion. Read and Chapelle (2001) therefore postulate that interactionalist approaches form the best construct descriptions for testing lexical knowledge, not least because they appear to fit with current communicative language teaching and testing paradigms. The question, however, is, to which extent vocabulary can be measured separate from the language skills. Views are split as to whether this can and indeed should be done. However, vocabulary measures based on trait definitions still seem to bear merit, least for diagnostic purposes. It therefore emerges that the intended purpose of a vocabulary test should determine its design and advantages and drawbacks of various options need to be evaluated to arrive at a sound decision which then has to be explicitly communicated.

2.2.3. Operationalizing the construct – key considerations in vocabulary testing

As hinted at in Section 2.1, the testing of lexical knowledge is, in some respects, considerably different from the testing of any other language skills. Several distinct features of vocabulary knowledge and vocabulary assessment have resulted in a tradition of vocabulary testing which is strongly characterized by "objective", psychometric approaches to assessment (Read, 2000). One of these features is the construct itself that appears to lend itself to (context-) independent, easily scorable test formats.

Words, or even phrases, are discrete, independent meaning units. Schmitt (2010) states that "vocabulary is largely item-based learning, and so each item addresses a separate construct" (p. 185). This clearly sets vocabulary knowledge apart from other language skills where the skills and sub-skill areas themselves, in as much as they are agreed to exist, are more interrelated. For instance, a certain degree of ability to read for specific details makes it likely that a candidate who does well on one item testing this reading behaviour can be expected to do well on a different item testing the same or a similar reading behaviour, given the text passages and tasks are of comparable difficulty. In terms of vocabulary, however, knowledge of one item does not necessarily imply the knowledge of another unit. In other words, a person's knowledge of the word *table* does not mean they are also familiar with the word *book*. A positive score on an item testing *table* has thus, theoretically and strictly speaking, no implication on the candidate's score on another item testing *book*.

In its most extreme form, this approach would entail that each word is its own construct. This, however, makes it not only an almost unmanageable abundance of constructs in terms of psychometric evaluations, but also severely limits the generalizability of any vocabulary test as any test score would have to be interpreted to provide information on the candidate's knowledge of that particular word tested and only that.

2.2.3.1. Item sampling - Word frequency

One way to overcome this is to assume a construct that clusters these discrete, independent meaning units in some form. Frequency, for instance, could be a clustering factor as we might hypothesize that learners are likely to learn the most frequent and thus useful words first so that some kind of relationship is underlying these units. This means that it is more probable that a learner who knows *book* (1K according to Nation, 2004) also knows *table* (1K according to Nation, 2004) than that they also know *audacity* (10K according to Nation, 2004). This broader construct aids not only statistical analysis, but more importantly a generalizable and meaningful score interpretation.

This approach was fostered by work into word frequency in the first half of the twentieth century, resulting in vocabulary lists for pedagogical purposes that provided "a large stock of vocabulary items that could be conveniently sampled to select the target words for a test" (Read, 2000, p. 76). Frequency might be a reasonable criterion to sample items and profile a person's knowledge as measured in a test. It has been shown to be a useful clustering factor and predictor of difficulty (Schmitt, 2010), particularly for high frequency bands and it helps circumvent to some extent the item-based psychometric and construct problems in vocabulary tests. However, frequency is not a sufficient predictor of knowledge (Schmitt, 2010). It seems a useful predictor of groups of words from a particular frequency band, but not necessarily for any individual word from that band. Also, frequency might be a less powerful clustering factor at lower frequency levels. For this reason, problems of internal consistency and equality of test forms are challenging, if not impossible, to resolve in vocabulary tests at this stage.

2.2.3.2. Item sampling - Unit of counting

Another key issue connected to the idea of frequency is that of the unit of counting when sampling vocabulary target items. Even leaving aside the issue of formulaic sequences and the fact that they are ubiquitous but not yet part of any systematic word list useful for sampling test items, it is still a matter of debate whether lemmas or word families should be the basis of vocabulary test sampling methods. Bauer and Nation (1993), as well as Nation and Webb (2011), claim that knowledge of one member of the word family implies that other members will also be known, at least receptively. Schmitt (2010) admits that there might be grounds for subscribing to this assumption when it comes to receptive word knowledge. The sampling of many existing vocabulary tests, such as the Vocabulary Levels Test (VLT) or the Vocabulary Size Test (VST), therefore rests on word family lists. However, the word family as best unit of counting has recently been contested (Schmitt, 2010; Schmitt & Zimmerman, 2002; Ward & Chuenjundaeng, 2009) and the psycholinguistic reality of word families is still undetermined (Schmitt, 2010). Even Nation (2016) recently acknowledged that Level 6 word families (Bauer & Nation, 1993) might be "too inclusive for lower proficiency learners of English as a foreign language" (p. 182). Aitchison (2003) showed that lemmas are much more reminiscent of the way our minds process vocabulary. The recent increase in lemmatized lists (Schmitt, 2010), for instance the new General Service List (Brezina & Gablasova, 2015) or the Essential Word List (Dang & Webb, 2016), could indicate that lemmas are gaining currency as a counting unit. Schmitt (2010) cites Nation that "for productive use, […] the lemma, or even word form, is the best unit of counting to use" (p. 192). Even though the unit of counting should be tailored to the purpose of the test or study, in the interest of comparability, the field might benefit from a standard unit that applies for both receptive and productive vocabulary tests, which would be a further argument for the use of lemmas (Schmitt, 2010). Lemmas also have the advantage of being transparent and exact in their definition, which is not necessarily the case with word families as different researchers have suggested different principles for word form inclusion (Schmitt, 2010).

Taking Sinclair's (2004) notion into account that different realisations of the same lemma might take very different collocations, even this unit of counting is not entirely unproblematic (Stubbs, 2009). However, sampling items from lists of word forms seems highly impractical due to the sheer amount of data. At the same time it needs to be acknowledged that different members of the same word family will indeed sometimes have very different characteristics, such as collocations. Using the lemma as counting unit could mean to steer a middle course. It limits the variability otherwise introduced by the word family but also introduces some clustering factor that renders sampling manageable. In any case, the selection of the counting unit is crucial for both the vocabulary test design and the validity and generalizability of score interpretations. Chapter 4 in this thesis will discuss this issue in more detail.

2.2.3.3. Cognates

The inclusion of cognates in vocabulary tests is also a topic of much debate. Elgort (2013), Gyllstad, Vilkaité and Schmitt (2015), and Laufer and McLean (2016) have demonstrated that the inclusion of cognates does have an effect on Vocabulary Size Test (VST) scores. Petrescu, Helms-Park and Dronjic (2017) recently attested cognate facilitation in VLT scores. However, Eyckmans et al. (2007) infer from their findings that removing cognates from a yes/no test does not improve or deteriorate the test's quality. While there needs to be an awareness by test score users that cognates might impact on scores, Nation and Webb (2011) follow Cobb's (2000) argument that the issue is less pressing for vocabulary knowledge tests than for vocabulary learning tests as the former type is not primarily concerned with how the learners have come to know a particular item. They therefore argue that excluding cognates and loanwords in vocabulary tests would confound the measure as it would not be entirely representative of a person's vocabulary size and would not allow for comparison of learners from different L1 backgrounds.

2.2.3.4. Translations

Given the unabated popularity of translation as a method of vocabulary pedagogy in many parts of the world, some vocabulary tests employ L1 translation (Barrow, Nakanishi, & Ishino, 1999; Snellings, van Gelderen, & de Glopper, 2004; Stalnaker & Kurath, 1935; Stubbe, 2013). Translation tests can come in different forms and formats. They might be bilingual versions of existing tests, such as the VST, which uses the multiple-choice format (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011). In other tests, candidates might literally be asked to provide a translation of a word or sentence or match a word or sentence with its respective L1 equivalent. Such translation formats have the advantage of being relatively easy to design and administer, but they are of limited use in international and increasingly multilingual teaching or research contexts. Also, Nation (1990) cautions that concepts in L1 and L2 might not always be identical, so that L1 and L2 words cannot by default be assumed to match exactly. This, in turn, might have adversary washback effects. Translation formats are therefore not very popular in vocabulary tests designed for larger scale research, although translation methods have been used in various research settings (e.g. Waring & Takaki, 2003) and can be useful in validation settings when verifying word knowledge.

2.2.3.5. Test formats – general issues

Schmitt (2010) claims that "[w]hen measuring knowledge of a lexical item, it is necessary to ensure that the test format does not limit the ability of participants to demonstrate whatever knowledge they have of the item" (p. 174). This, however, has to be questioned. Although the incremental nature of vocabulary learning certainly implies that vocabulary tests should account to some extent for partial, developing knowledge of a word, Schmitt's principle needs to be problematized in terms of score interpretation and test purpose. The choice of format certainly depends on the kind of information and the degree of precision of knowledge a test developer or user is aiming for.

However, there is definitely agreement that the format and instructions of a test should not introduce construct-irrelevant difficulties (Schmitt, 2010). One generally reasonable way to aim for this is by using defining vocabulary which is of the highest possible frequency level or at least of higher-frequency than the target item. Frequency counts, however, should be corroborated using different source corpora (Schmitt, Dörnyei, Adolphs, & Durow, 2004). Care needs to be taken for these frequency restrictions for definitions not to result in unnatural or contrived formulations (Schmitt, 2010). This, however, has been implemented in existing vocabulary tests with varying degrees of success.

As abovementioned, short, discrete, context-independent, selected response formats have been, and continue to be, popular "objective" means of measuring vocabulary knowledge. Vocabulary's aptness for objective, discrete-point testing, combined with influential advances in psychometrics, means that multiple-choice items were and continue to be the most popular response format to test vocabulary knowledge. Lado (1961), as early as the early 1960s, observed that "[t]he multiple-choice type of item has probably achieved its most spectacular success in vocabulary tests" (p. 188). Nation's VST constitutes probably the most prominent recent example of this.

This is at least somewhat surprising given the criticisms and precautions that have legitimately been voiced against this format. Nation and Webb (2011) maintain that taking multiple-choice vocabulary items is not like normal language use where we do not encounter meaning choices for (unknown) words. An early study by Goodrich (1977) indicated unsurprisingly that the nature of the distractors has a considerable impact on the measurement and its outcome. Wesche and Paribakht (1996) claim that the meaningfulness of multiple-choice question (MCQ) scores is limited as test takers might arrive at the correct answer by guessing or a process of elimination, testing their knowledge of distractors as much as their knowledge of the target word. They also raise concerns about how this format deals with polysemous targets and report difficulties pertaining to the construction of functioning items of this format. In this way, the format faces similar problems for testing vocabulary knowledge as it does for testing other language areas and skills. However, other points Wesche and Paribakht (1996) problematize with this format, such as limited sampling rates, do not seem to outweigh the benefits of practicality. Particularly so, as practicality in terms of test administration and scoring is one of the key advantages of MCQ items, which probably allow for higher sampling rates than most alternative test formats, including Wesche and Paribakht's (1996) Vocabulary Knowledge Scale (VKS). Recently, Stewart (2014) investigated whether MCQ items inflated test scores due to guessing and argued that multiple choice items generally overestimate vocabulary size. However, his argument is based on findings of a simulation study rather than real test taker data, and has not remained uncontested (Holster & Lake, 2016; Stewart, McLean, & Kramer, 2017). It is clear, however, that our understanding of the workings of MCQ items, in particular pertaining to their difficulty and their proneness to guessing, is currently insufficient given the popular use of the format.

The disappointingly low number of studies investigating the usefulness of MCQ items for vocabulary knowledge measurement might be due to the fact that there are generally very few validation studies available in vocabulary assessment research overall. One of the reasons for this could be the difficulty of validating vocabulary tests due to their specific nature. Correlational validation studies, frequently employed in language testing, have been attempted (e.g. Fitzpatrick & Clenton, 2010; Sims, 1929; Tilley, 1936), but are in principle devoid of a generally accepted standard vocabulary knowledge measurement instrument to compare other tests with.

Also, the ongoing debate in the language testing community about validity theories and validation frameworks, which is in itself far from settled, seems to disregard vocabulary assessment. This is perhaps due to the overwhelming impact of the communicative approach in language testing, resulting in a lack of validation frameworks and models for vocabulary tests. While some seem at least adaptable for this purpose (Kane, 2006), others, such as Weir's (2005) socio-cognitive validation approach, published at length also in terms of the more traditional language skills (Geranpayeh & Taylor, 2013; Khalifa & Weir, 2009; Shaw & Weir, 2007; Taylor, 2011), do not even account for the possibility of vocabulary test validation.

Another format for vocabulary assessment popularized in the 1970s is the cloze test (Read, 2000). In an attempt to introduce contextualization into vocabulary tests while at the same time retaining some control over the tested lexical target items, this format required candidates to fill predetermined gaps at fixed ratios (at every n-th word) in written texts. Several variations of the format have been suggested, such as the rational cloze (with principled selective word deletion instead of according to a given ratio), multiple-choice cloze formats or the C-test (deleting the second half of every second word) (Klein-Braley & Raatz, 1984). Although not primarily intended as sole measures of vocabulary knowledge, the assumption that the completion of these tasks required test takers to draw heavily on their lexical resources made them attractive for vocabulary researchers (Chapelle, 1994). Singleton (1999) even claims that "C-test data are essentially lexical data" (p. 205).

The major problem with cloze tests of any form and the reason they have by now generally fallen out of favour for vocabulary assessment, however, is exactly that assumption. In fact, researchers are still uncertain what it is precisely that these tests are measuring (Bachman, 1985; Eckes & Grotjahn, 2006; Jonz, 1990). Lexical knowledge almost certainly plays a role in answering such items, but it seems also most likely that reading skills and other areas of linguistic knowledge form at least part of their construct (Alderson, 1979; Eckes & Grotjahn, 2006; Porter, 1983) and render it less useful as a distinct vocabulary measure. While cloze procedures have been used for various purposes and eventually just vaguely suggested to test 'overall language proficiency', vocabulary researchers, though acknowledging they might be indicative of a person's lexical ability to some extent, have now discarded them as feasible means of testing (exclusively) vocabulary knowledge. Read (2000) trenchantly concludes that "a cloze test tends to make a very *embedded* assessment of vocabulary, to the extent that it is difficult to unearth the distinctive contribution that vocabulary makes to test performance" (p. 115, emphasis in original).

Although several studies (e.g. Arnaud, 1989; Corrigan & Upshur, 1982) and particularly findings from corpus linguistic research (Römer, 2009) "challenge the notion that vocabulary can be assessed as something separate from other components of language knowledge" (Read, 2000, p. 115), this embeddedness is a serious threat to construct validity if the test result is first and foremost taken as an indication of a person's lexical knowledge. The value of cloze tasks seemed to lie in the contextualization they provided for the target items that went beyond traditional vocabulary testing methods and thus the provision they made for adopting a broader view of vocabulary that included multi-word phrases, idioms and other formulaic units. However, this amount of contextualization is not without problems as more than lexical knowledge could be tested in such tasks (Read, 2000).

The problem with the threshold at which embeddedness becomes potentially construct-irrelevant to the measurement (Weir, 1990) holds also for vocabulary assessments that focus on indices of lexical richness or lexical sophistication in written or spoken learner productions. There are a number of indirect vocabulary assessments, in which information about a person's lexical knowledge is gleaned from pieces of speech or writing: Lexical frequency profiles (Laufer & Nation, 1995), type-token ratios (Arnaud, 1984; Daller & Phelan, 2007; Daller & Xue, 2007; van Hout & Vermeer, 2007) or adjusted similar indices such as Guiraud's Index, Advanced Guiraud (Daller, van Hout, & Treffers-Daller, 2003), D (Malvern & Richards, 1997), Measure of Lexical Richness (Vermeer, 2004), Coh-Metrix (Crossley, Salsbury, McNamara, & Jarvis, 2011), Limiting Relative Diversity (Malvern, Richards, Chipere, & Durán, 2004) or P-Lex (Meara & Bell, 2001). However, all of these suffer from the validity issue that it is almost impossible to disentangle vocabulary knowledge from writing or speaking skills in such measurements. The information they provide can be highly valuable to complement the picture of a language learner's status or progress in an integrative manner, but it seems limited as the sole source for establishing someone's lexical knowledge.

2.3.Existing vocabulary tests - An evaluation

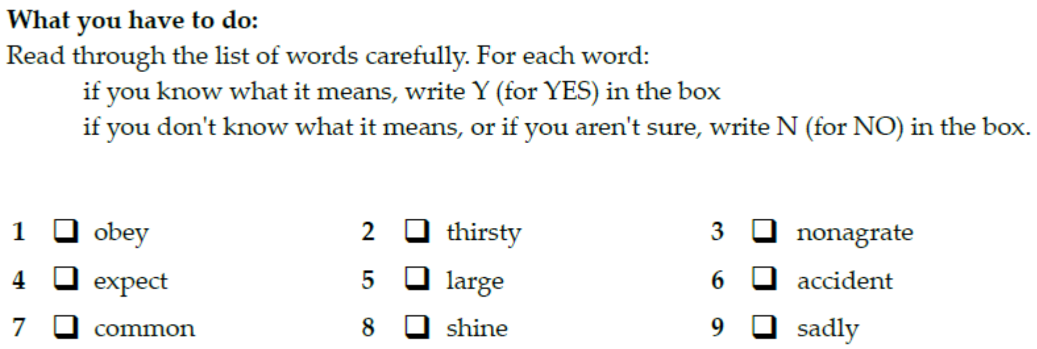
To date, there is no standardized and unequivocally accepted measure of vocabulary knowledge available that is backed and validated by empirically sound findings (Schmitt, 1999). This is despite or probably because a large number of different tests have been developed and used for an array of diverse purposes.

2.3.1. Measurements of breadth of vocabulary knowledge

Nation and Webb (2011) identify three main approaches to assessing the breadth of a learner's vocabulary knowledge: "(1) counting the words that someone produces, (2) counting the number of words in a dictionary and testing what proportion of these are known, and (3) sampling from various frequency levels and testing to estimate the amount of vocabulary known at each level" (p. 196). A number of researchers subscribe to the first approach and have developed sophisticated measures to assess the lexical richness of learner production (see Section 2.2.3.5). These measures all suffer from the disadvantage that learners do not produce all the words they know in any performance so that they might not be adequate assessments of vocabulary breadth. Only a few studies have approached the issue of vocabulary breadth measurement by sampling from dictionaries. The study by Goulden, Nation and Read (1990) constitutes an exception to this, but their resulting test was a self-assessment tool rather than a validated test (Read, 2013). Generally speaking, most vocabulary tests have relied on the third approach, which seems the most useful in terms of construct validity and generalizability. The most prominent of these tests will be evaluated in the following sections.

2.3.1.1. Yes/No Checklist tests

First suggested by Meara and Buxton (1987) as an alternative to established multiple choice tests of L2 vocabulary knowledge in terms of breadth, these tests intend "to measure learners' receptive vocabulary size by presenting them with a sample of words in the target language covering certain frequency levels and asking them to indicate the words they know the meaning of" (Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001, p. 236). In other words, test takers simply check against a list of target items whether they know it (Yes) or not (No), hence the name of the instrument. Meara (1994) states that this test measures the most basic of word skills, i.e. "*the* basic skill on which all other skills depend" (p. 6). An example item can be seen in Figure 4.



What you have to do:

Read through the list of words carefully. For each word:

if you know what it means, write Y (for YES) in the box

if you don't know what it means, or if you aren't sure, write N (for NO) in the box.

1 □ obey

2 □ thirsty

3 □ nonagrate

4 □ expect

5 □ large

6 □ accident

7 □ common

8 □ shine

9 □ sadly

Figure 4: Example items from a yes/no checklist test (Meara, 1992, p. 18)

A percentage of “pseudowords” (Beeckmans et al., 2001), “non-words” (Read, 2007) or “imaginary words” (Meara & Buxton, 1987), varying in ratio in different studies (Abels, 1994; Hacquebord, 1999; Meara, 1992; Meara & Buxton, 1987), is generally added to the list to counter learners’ overestimation of their vocabulary knowledge.

Perhaps the most severe weakness of the instrument, particularly when used for research purposes, appears to be the fact that at no point the actual knowledge of one or multiple meanings of a target word is verified (Eyckmans et al., 2007). The tool therefore essentially remains a self-assessment instrument. This might lead to students overestimating (Mochida & Harrington, 2006) or, in rarer cases, underestimating (Stubbe, Stewart, & Pritchard, 2010) their vocabulary knowledge. Because candidates are not required to demonstrate knowledge of the word at any stage, Pellicer-Sánchez and Schmitt (2012) found learner overestimation even with participants that had not checked any non-words. Shillaw (1999) found that the Japanese learners in his study were conservative in their estimates and very rarely checked non-words at all. In addition to difficulties in score adjustment discussed below, these findings render the entire non-word approach questionable.

The format, however, has also been criticized for not permitting the testing of multiple word meanings (Abels, 1994) or of multiple dimensions of word knowledge (Beeckmans et al., 2001), thereby curtailing the view of vocabulary knowledge to a very simplistic notion. Further, the optimal length of the test for representative sampling and adequate size estimates has not been corroborated (Beeckmans et al., 2001), neither have standardized guidelines for the construction of pseudowords been empirically validated (Beeckmans et al., 2001). Also, there is no research that investigates whether more specific instructions had any effect on candidate scores. The test has also been found not to yield reliable results with low-level learners (Meara, 1996), which appears problematic, least from a diagnostic perspective.

A major advantage of the format, employed for instance in the Eurocentres Vocabulary Size Test (Meara & Jones, 1990), the DIALANG Vocabulary Size Test (Alderson, 2005), and the massive online experiment by researchers from Ghent University (Keuleers, Stevens, Mandera, & Brysbaert, 2015), lies in its ostensible practicality and its potential to test a large sample of words in relatively short time (Mochida & Harrington, 2006). However, although it seems easy and quick to construct, administer and correct a vocabulary test using this response format, scoring the answers of test takers is disproportionately tricky and complex and the interpretation of the scores even more so.

The scoring problem mainly stems from the four possible combinations of kinds of items and potential responses. Candidates may (1) tick that they know a real word (true hit), or (2) tick a pseudoword (false alarm), or (3) not tick a real word (miss), or (4) not tick a pseudoword (correct rejection). This allows for several scoring methods. Beeckmans et al. (2001) report that the most straightforward scoring procedure, i.e. simply adding the correct responses, has rarely been considered as researchers seem to agree that the false alarm rate has a significant role to play to correct for overestimation. Huibregtse, Admiraal and Meara (2002) even maintain that "hits" and "correct rejections" cannot be considered equivalent as they are acceptable "in different ways and for different reasons" (p. 231).

In attempts to salvage the undoubtedly attractive format in terms of practicality and administrability, various researchers have suggested different scoring procedures and sophisticated correction formulae over the past years.

Meara (Meara, 1992) suggested a scoring formula Δ_m based on Signal Detection Theory (SDT). This, however, has been exposed as problematic because an individual's response style, though clearly a factor outside the scope of the construct purported to be measured, appears to have a major influence on the test score. Eyckmans et al. (2007) claim that "[w]hen in doubt, the testee may lean towards either a *Yes* or a *No* response simply because of his response style (overestimation, underestimating, refraining from answering) or attitude (analogous to Bourdieu's 'economy of practice')" (p. 62). Thus, "small differences in response behavior [individual response style] may cause large differences in scores" (Huibregtse et al., 2002, p. 229) because the score rapidly approaches 0 for moderate performances "even if the performance is well above chance level" (Huibregtse et al., 2002, p. 229). What is more, this "presence of a response bias artificially enhances the reliability of test data" (Eyckmans et al., 2007, p. 63; Eyckmans, 2004), resulting in a risk for test users to place "too much confidence in tests which, even though reliable, actually measure a different construct than the one aimed for" (Eyckmans et al., 2007) and thus lack validity (Beeckmans et al., 2001; Eyckmans, 2004). Controlling for this response behavior has proven to be difficult, both with the use of correction formulae and with more controlled computer interfaces for test delivery (Eyckmans et al., 2007).

Other researchers have proposed models for correcting scores for (blind) guessing behavior. These formulae, however, have been problematized as they also fail to account for individual response styles (Huibregtse et al., 2002). Even after positing the I_{SDT} formula, which neutralizes response style differences and corrects for guessing (Huibregtse & Admiraal, 1999), Huibregtse, Admiraal and Meara (2002) still conclude that [t]he question of what would be an appropriate interpretation of the test score remains" (p. 242).

The issue could also not be resolved by an innovative psycholinguistically-motivated approach put forward by Pellicer-Sánchez and Schmitt (2012), in which scores were combined with reaction times for correction purposes. They conclude from their study comparing scores obtained by native and non-native speakers of English taking the test with the candidates' actual word knowledge as elicited in personal follow-up interviews that "there was no clear advantage for any of the [correction] approaches under comparison, but their effectiveness depended on factors like the false alarm rate and the size of participants' overestimation of their lexical knowledge" (Pellicer-Sánchez & Schmitt, 2012, p. 489). This reiterates Beeckmans et al.'s (2001) claim that the "Yes/No format in its current form does not meet the required standards in terms of reliability" (p. 272) and "suffers from a bias which cannot be handled by one of the correction methods while maintaining a sufficiently accurate measurement" (p. 272). Stubbe's (2013) recently suggested regression formula (see also Stubbe & Stewart, 2012) seems to work better than existing correction formulae, but it remains to be demonstrated how this adjustment works in various settings and with different and differing populations (In'nami, 2013).

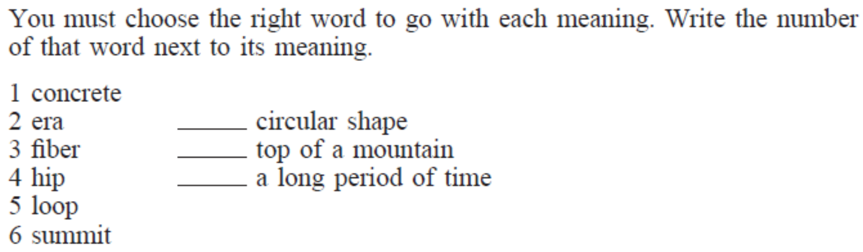
Meaningful score interpretation of checklist scores, by extension, is incredibly challenging as it seems unclear, even for true hits, whether one meaning is known or several meanings are mastered. Some variations of this format also neglect the intention of it being a forced choice test (Beeckmans et al., 2001), which renders score interpretation even more difficult.

This obviously limits the usefulness of this checklist format for vocabulary knowledge measurement purposes. While Read (2007) maintains that "[d]espite its simplicity, the Yes/No format has proved to be an informative and cost-effective means of assessing the state of a learner's vocabulary knowledge, particularly for placement and diagnostic purposes" (p. 112-3), it could be argued that the format is rather unfit, particularly for diagnostic purposes, as it is such a coarse measure of vocabulary size with very little information content for score users. Eyckmans et al. (2007) conclude that "[t]he Yes/No format is too susceptible to the interference of construct external variables" (p. 75), which poses a considerable threat to any interferences drawn from its scores (Fulcher, 2003).

2.3.1.2. Vocabulary Levels Test (VLT)

The Vocabulary Levels Test (VLT) is arguably the nearest thing to a standardized vocabulary test currently available (Meara, 1994, 1996; Schmitt, Schmitt, & Clapham, 2001). Designed initially as a diagnostic tool for teachers (Nation, 1983, 1990), it has come to be used as a widely employed instrument amongst teachers and researchers alike to provide an estimate of vocabulary breadth (and often inappropriately of vocabulary size) of L2 language learners (Cobb, 1997; Laufer & Paribakht, 1998; Read, 1988; Schmitt & Meara, 1997; Shiotsu & Weir, 2007).

The practicality of the test thereby seems to have been a driving force of this development. Using a multiple matching format, test takers are presented with six words in a column on the left and the corresponding meaning senses of three of these in another column on the right. They are then asked to indicate for each meaning sense in the right-hand column which single word from the left-hand column it matches. An example item can be seen below.



You must choose the right word to go with each meaning. Write the number of that word next to its meaning.

1 concrete
2 era ______ circular shape
3 fiber ______ top of a mountain
4 hip ______ a long period of time
5 loop
6 summit

Figure 5: VLT example item (Schmitt et al., 2001, p. 58)

Each cluster thereby targets three words, although some researchers have argued that knowledge of the meaning of the three distracter words is also tested as the test takers need to be familiar with them when they discard them (Read, 1988). There is a fixed number of clusters for each frequency level from which the target words were sampled, hence the name "levels test". Within each level, the sample is stratified to represent the distribution of English word classes. This ratio is either 5 (noun) : 3 (verb) : 1 (adjective) (Beglar & Hunt, 1999) or 3 (noun) : 2 (verb) : 1 (adjective) (Schmitt et al., 2001). Word classes are not mixed within any one cluster. The clusters are sampled in equal amounts from the 2K, 3K, 5K and 10K frequency bands of word family lists and the University Word List (UWL) (Xue & Nation, 1984) or, more recently, the Academic Word List (AWL) (Coxhead, 2000).

Cameron (2002), in a direct comparison of the VLT and Meara's (1992) yes/no test, found that the VLT was a more useful tool to profile learners' vocabulary knowledge. She also reports a relatively atypical profile across frequency levels. However, her study is limited due to the use of an outdated 18-item version of the VLT and a very small sample population size. Nevertheless, her findings must be credited for being one of the few investigations of the format's usefulness involving secondary school students rather than participants from tertiary education contexts.

In light of the popularity of the VLT, it comes as a surprise that only very few studies have investigated the validity of the instrument. One of the few validation studies conducted revealed that an implicational scale can be assumed for the frequency levels (Read, 1988). Candidates who knew lower-frequency words usually also knew high-frequency words. Beglar and Hunt's study (1999) focused on the validity of the 2K and the University Word List sections of the VLT. They found that, psychometrically speaking, these sections of the VLT assessed a single construct. However, they also claimed that item difficulty needed to be explored more thoroughly in further validation studies as well as potential item interdependence in the chosen matching format. Their study also raised concerns about the representativeness of the sampling of early versions of the VLT. Beglar and Hunt (1999) further voiced apprehensions about the interpretation of VLT scores, the sampling of which is based on word family frequency lists. They state that "knowledge of a word's base form does not guarantee knowledge of its derivatives or inflections" (Beglar & Hunt, 1999, p. 147). Beglar and Hunt's study, however, suffers from severe limitations due to its narrow population of learners from only one L1 background.

An attempt to overcome some of these shortcomings of the VLT was undertaken by Schmitt, Schmitt and Clapham (2001). They set out to deliver a more comprehensive validation study of newly designed VLT versions based on the weaknesses identified by their own and earlier research. In the construction of the revised test versions, they adhered to most of the original VLT design principles (Schmitt et al., 2001).

- Candidates are asked to recognize the form rather than the meaning, i.e. the options are words instead of definitions.
- Definitions are deliberately kept easy, using only words from the same or higher frequency levels, and short so as to keep reading to a minimum and not to muddy the measurement (Weir, 1990).
- The format thereby accounts for the incremental nature of vocabulary knowledge by tapping into partial word knowledge.
- One feature that helps this is the design of the clusters, which contain semantically and orthographically very distinct options.
- Within the clusters, target words options are ordered alphabetically and definitions are ordered according to length to reduce guessing.
- The target words are presented in their most frequent form of the word family, which in most cases is the base form.
- In cases of derivatives, Level 5 of Bauer and Nation's (1993) model was selected as the cut-off point of admissible forms.

Schmitt, Schmitt and Clapham (2001), however, not only provide a detailed account of the rationales, principles and extensive trialing that guided the test revision, but also offer a thorough validation study with 801 EFL learners from different countries, comprising item analysis, profile analysis, factor analysis, reliability and equivalence analysis as well as an investigation of the "concurrent validity of the tests by correlating the results with the results of an interview" (p. 57) with 22 candidates on a third of the tested items.

They conclude that items perform reasonably independently of each other, that "discrimination indices for the Levels Test are acceptable, bearing in mind that vocabulary is learned as individual units" (Schmitt et al., 2001, p. 66) and that guessing does not seem to be a serious threat to the validity of the scores. The frequency sections further allowed for implicational scaling as facility values decreased as a function of the frequency level. They also asserted that the VLT globally seems to be measuring a unidimensional trait, but question the usefulness of factor analytic approaches in this respect since at individual word level “the only construct which makes any sense is ‘knowledge of that particular word’s properties’” (Schmitt et al., 2001, p. 71). Instead, they suggest to cluster hypothesized factors according to frequency levels.

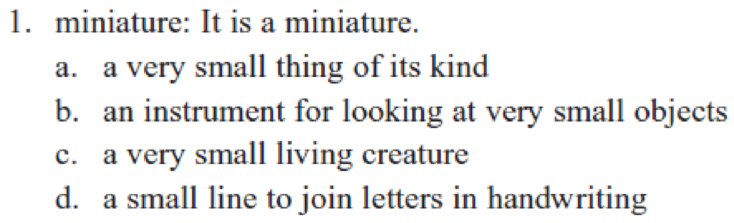
Their test versions with an increased 30 items per level compared to previous test versions yielded good reliability values while not compromising the practicality of the instrument. Findings from the interviews suggested that the VLT, despite being a selected response format, showed relatively few problems with guessing behavior distorting results, thus reflecting underlying, even if only partial, lexical knowledge. Their proposed revised versions were also able to generate similar, “if not truly equivalent, scores” (Schmitt et al., 2001, p. 79), which is why these latest versions are now the most widely used. Xing and Fulcher (2007), however, caution in their reliability assessment of these versions at the 5,000 word frequency level that they may not be regarded as parallel forms.

Schmitt, Schmitt & Clapham (2001) acknowledge that guessing and item interdependence might be problems of this test format that would require further investigation. Kamimoto (2008) and Webb (2008) suggested there was a 17% chance of learners blind guessing correct responses. Stewart and White (2011) maintain that the issue of guessing is further complicated in the VLT format as distractors are words chosen from the same frequency level as the targets, i.e. from the tested domain. This means that the overestimation in scores due to guessing is variable depending on the proportion of distractors known to a candidate. Therefore, the probability of a successful guess is, in this format, not simply a function of the number of distractors used. Stewart and White (2011) ran multiple guessing simulations on the VST and found that candidates’ scores are generally and consistently inflated by 16-17 points on a 99-item VLT test “until over 60% of words are known, at which point the score increase due to guessing gradually begins to diminish” (p. 378).

Like most other validation studies of vocabulary measures, Schmitt, Schmitt & Clapham's (2001) validation study also mostly employed intermediate to advanced language learners at university level which leaves open questions about the validity of scores generated with the VLT when used with lower-proficiency or younger EFL learners. Xing and Fulcher (2007) further note that the word lists, on which the test versions are based, must by now be considered out of date and therefore highlight the need for an updated measure of vocabulary breadth. Xing and Fulcher (2007) also point out that Schmitt's suggested cut-off of 80% correct answers for a level to be considered acquired, needs to be empirically asserted. Most importantly, though, Schmitt, Schmitt and Clapham (2001) explicitly state the purpose of the VLT and outline what it may or may not be used for appropriately and which claims could be regarded valid and which not. This must be acknowledged as it is a practice surprisingly seldom implemented in vocabulary assessment development and research.

2.3.1.3. Vocabulary Size Test (VST)

A different test of vocabulary breadth gaining increasing popularity in second language acquisition (SLA) research is the Vocabulary Size Test (VST) (Nation, 2008; Nation & Beglar, 2007; Nation & Gu, 2007). This fairly recent instrument claims to establish a total estimate of written, receptive vocabulary size. The test uses a four-option multiple choice format in which candidates are presented with a target word, a short, non-defining sentence in which the target word occurs in bolded print, and four alternative definitions of the word in question, one of which is the key. An example item is displayed in Figure 6 below.



1. miniature: It is a miniature.
   a. a very small thing of its kind
   b. an instrument for looking at very small objects
   c. a very small living creature
   d. a small line to join letters in handwriting

Figure 6: VST example item (Beglar, 2010, p. 104)

The multiple choice format makes the test very practical in terms of administration and scoring. 10 target words per 1K band are thereby selected from the 14,000 most frequent word families of English according to the BNC corpus. More recent versions, using only 5 items per 1K band, but extending the test to the first 20,000 most frequent word families in the BNC, have recently been made available (Nation, 2014), even though "[i]t is difficult to conceive of contexts in which it would be necessary to measure the written receptive lexical knowledge of second language learners of English beyond the 14,000-word frequency level" (Beglar, 2010, p.116).

The construction, validity and usefulness of these versions need to be questioned as even for the original 14K version of the VST, validity evidence is sparse and rather mixed. A total vocabulary size estimate is arrived at by multiplying the scores by 100 or 200 respectively, each test item thus representing 100 or even 200 word families. Gyllstad (2012), in a classical test theory approach to validating the VST, states that the VST shows promise in yielding reliable scores, but that some items require revision.

Beglar (2010) offers a tentative Item Response Theory (IRT) validation attempt of the VST using a common item design with internal anchors and concludes that the majority of VST items show adequate fit to the Rasch model and contribute strongly to the hypothesized psychometric unidimensionality underlying the test. According to his findings, the items yield high measurement invariance in various test forms and allow for a precise measurement of candidate's ability with low standard errors and high reliability estimates.

Examining several aspects of validity, Beglar (2010) further concludes that "10 items per level is more than sufficient to estimate the test takers' lexical knowledge with a high degree of precision" (p. 107), buttressing the representativeness of the sampling in the VST. However, in the same publication Beglar (2010) notices the potentially problematic effects of such low sampling rates as he observes that one particular level was easier than expected "in part due to one extremely easy item" (p. 109). The sampling rate seems even more obscure when considering the rate employed per frequency level in the VLT (Schmitt et al., 2001).

Indeed, Gyllstad, Vilkaité and Schmitt (2015) found in their study comparing VST scores with follow-up interviews on 100 items, that 10 items might not be a sufficient sampling rate to represent any level. Instead they suggest a rate of about 30 items per 1,000 word family frequency level. In light of these findings, Beglar's (2010) claim that "it is possible to substantially reduce the number of items per word frequency level without encountering significant reductions in measurement precision" (p. 107) needs to be called into question as much as his claim that the 140 item VST version is responsive and sensitive enough to changes in vocabulary size for it to be an adequate SLA research instrument. The interpretability of VST scores, though briefly addressed in Beglar's article, may not emerge from an IRT study like the one he conducted and needs to be problematized given Gyllstad, Vilkaité and Schmitt's (2015) findings about the VST's proneness to guessing.

Zhang's (2013) findings suggest that the inclusion of an "I don't know" option and a penalty instruction reduce the amount of guessing in the VST. Comparing three versions of the original VST, one unchanged (Version 1), one with an "I don't know" option (Version 2) and one with both this option and a penalty instruction (Version 3), he found that the versions with the additional option were completed faster, yielded slightly better reliability indices and were better at separating learners according to vocabulary size in terms of Rasch separation and discrimination indices. In addition, for the group taking the original VST, the total raw scores were highest and significantly different from the average scores of the other two test version groups. Relating the VST scores to scores on a meaning recall task, however, he concludes that in terms of verified word knowledge, the groups' scores were not significantly different according to their actual vocabulary size. The comparisons also showed that guesses were common in all test versions, but were significantly less frequent in Versions 2 and 3. The additional option and the penalty instruction therefore seem to successfully discourage guessing to some extent. However, Zhang (2013) argues that they also discourage partial knowledge which the

VST explicitly intends to measure as well. Including an “I don’t know” option therefore seems to be a trade-off between reducing the number of random successful guesses and educated successful guesses guided by partial knowledge. The decision to include it thus depends on the test purpose and the precision of the word knowledge required.

Another problematic aspect of the VST is its basic assumption that “language learners beyond a beginning proficiency level have some control of word building devices and are able to identify both formal and meaning-based relationships between regularly affixed members of a word family” (Beglar, 2010, p. 103), thus justifying the word family at Level 6 of Bauer and Nation’s (1993) scale of levels as a counting unit for the sampling of test items. The hypothesis rests on evidence from studies which have identified word families as psycholinguistically real unit (Bertram, Baayen, & Schreuder, 2000; Bertram, Laine, & Virkkala, 2000; Nagy, Anderson, Schommer, Scott, & Stallman, 1989). This, however, is not universally agreed (Schmitt, 2010).

Also, research has also shown that derivational forms are problematic for language learners at various levels (Schmitt & Zimmerman, 2002; Ward & Chuenjundaeng, 2009), which relativizes this underlying assumption and could be taken as an argument against the selected counting unit in favor of lemmas. Also, while the assumption may hold true to some extent for receptive word knowledge used in reading and listening, it certainly falls apart when learners are asked to produce derivational forms (Schmitt, 1999; Schmitt & Zimmerman, 2002). Ward and Chuenjundaeng (2009) argue based on their results from a translation test that even in receptive word knowledge, “the use of word families as a counting tool leads to highly misleading conclusions” (p. 461) about learners’ vocabulary size. Scores obtained with a measure based on this notion could therefore overestimate what learners actually know and can do with representatives of word families and related word family members. Also, it has yet to be empirically shown at what point this alleged proficiency threshold lies beyond which mastery of regular affixation can be taken as a given.

2.3.1.4. Computer-Adaptive Test of Size and Strength (CATSS)

Laufer et al. (2004) make a strong case for including the strength of a candidate's vocabulary knowledge in lexical knowledge measures. They argue that, within a subcomponent such as form-meaning link, there are differences in how well a word is known and hypothesize a continuum of strength of word meaning knowledge ranging from form recall (Schmitt, 2010), what they term active recall, to meaning recognition (Schmitt, 2010), or passive recognition (Laufer et al., 2004). This dimension of strength is independent of the quality of the depth of word knowledge in terms of how well different aspects of word knowledge are mastered (collocations, associations, register, etc.). In their development of a monolingual and a bilingual Computer-Adaptive Test of Size and Strength, they set out to examine this implicational scale of difficulty and empirically tested whether recalling the form for a given concept was indeed more challenging than recalling a meaning for a given L2 word form, and whether this, in turn, was again more challenging than merely supplying the L2 form of a concept or providing the meaning of an L2 form in a recognition task as tested via four-option multiple choice items. Their test construction was based on three core assumptions (Laufer et al., 2004; Laufer & Goldstein, 2004):

1) That the ability to establish the link between word form and word meaning was the most important component of word knowledge.
2) That knowledge of the form-meaning link is incremental rather than an all-or-nothing phenomenon and that there is an underlying implicational scale of degrees of strength of word knowledge as described above.
3) That mastery of many words and their most frequent meaning sense is more important than mastery of a few words in depth, which, by that logic, highlights the rationale of the test design to focus on vocabulary meaning size in vocabulary assessment.

Their study preparing the construction of the CATSS is probably most valuable for establishing and corroborating this implicational scale from form recall to meaning recall to form and/or meaning recognition. The resulting instrument presents the candidate with definitions of target words sampled from different frequency bands (30 per 2K, 3K, 5K, 10K and AWL band) and test items asking to provide the form, i.e. recall or activate the highest level of word meaning knowledge. At this highest stage the initial letter of the target word is provided to guide the candidate and keep the acceptable options narrow. At the level of meaning recall the target word is presented in a sentence with a gap into which the candidate is required to insert the meaning of the target word. CATSS is computer-adaptive in that it presents the candidate with items asking for the second highest form of word meaning knowledge for those target words that have not been answered correctly in the first attempt. If an item is answered correctly at any of the four stages, the candidate is not exposed to it again. If an item is answered incorrectly at any of the higher stages, the candidate is presented with the item again in the subsequent stage until the answer provided is correct or the item is answered incorrectly at the fourth and lowest stage of word meaning knowledge.

The test, however, suffers from a number of practical problems, apart from sloppy typography in some items. In a number of cases it is not clear for the candidate which derivative form is required and the test would not handle alternative but theoretically acceptable forms. Neither will plural forms be accepted at the form recall stage, although they would also demonstrate the candidate's knowledge of the word. The non-acceptance of possible alternative answers becomes even more salient in the meaning recall stage, which inadvertently allows for a vast number of correct answers that should be accounted for. Laufer et al. (2004) maintain that "[o]n the basis of extensive piloting, the most frequent correct responses are included in the key" (p. 211). However, this extensive piloting procedure and the resulting key are not specified in more detail and even low-frequency answers would have to be accepted if they are found to demonstrate knowledge of the word meaning. A problem with higher-level learners might be that they know synonyms for target words which start with the same letter as the target word. This issue of accepting valid but non-target answers is not addressed by the test developers. Another weakness of the CATSS is that the target words are sometimes presented in different derivative versions of a word family at different stages, which might pose problems for candidates. Furthermore, it remains unclear whether the meaning recall stage, where, for instance, a synonym needs to be provided by the testee, means that it is the target word and not the knowledge of the synonym being actually tested.

While it thus appears that the basic assumptions underlying the CATSS are indeed useful and potentially valid, the operationalization of the parameters leaves room for improvement. The potentially most beneficial finding of the studies leading up to the design of the CATSS might be the established scalability of this strength continuum of word meaning knowledge in both the monolingual and the bilingual version of the test. However, Laufer et al.'s study (2004) has to be interpreted with some caution as it could be regarded to feature a methodological flaw in that all of the participants were of relatively high proficiency and none of their participants actually took the four levels of the test to establish the relative difficulty of the items.

2.3.1.5. Picture-based vocabulary tests

According to Nation (1990), "[i]n recognition tests, we want to see if the learners know the meaning of a word after they hear or see it. In such tests the learners hear or see an English word and then […] (c) choose one from a set of pictures, mother-tongue words, or English synonyms or definitions" (p. 79-80). In order to address concerns about keeping the involvement of linguistics skills and knowledge other than vocabulary knowledge to a minimum, some test designers resort to using pictorial prompts or inputs in their lexical assessment tools.

The use of pictures in language testing is not without risk as they may be culturally loaded and ambiguous and thus jeopardizing positive interactiveness of a test (Bachman, 1990). If used appropriately, however, they provide an invaluable source, particularly for vocabulary assessment, where they could make lengthy definitions redundant. This would not only decrease any risks of involving reading skills, syntactic knowledge or other construct-irrelevant aspects but would also potentially facilitate the complex task of the test designer to contrive definitions that contain words of a lower frequency than the actual target.

One example of such a picture-based vocabulary test is the Peabody Picture Vocabulary Test (PPVT) (Dunn & Dunn, 1959). It is particularly prominent in L1 research (Siyanova-Chanturia & Martinez, 2014), but has also been used in several L2 research settings (e.g. Sparks et al., 1998; Tomiyama, 2008; Unsworth, Persson, Prins, & De Bot, 2014).

Now in its fourth edition, the PPVT-IV (Dunn & Dunn, 2007) measures word knowledge by means of multiple-choice items, in which four full-colour pictures make up the response options per target word. The candidate is asked to "select the picture that best illustrates the definition of the word" (Hoffman, Templin, & Rice, 2012, p. 754), which is presented only in oral form to the candidate by a test administrator. The 228 test items in total cover 20 content categories "(e.g., actions, vegetables, tools)" (Pearson, 2014) and the three parts of speech nouns, verbs and adjectives. The items are sampled from reference works according to categories and norm-referenced (age or grade) difficulty level instead of frequency. The test designers assure that all "[i]tems were reviewed and empirically analyzed for difficulty, validity (discrimination), and freedom from bias with respect to sex, ethnicity, geographic region, and SES" (Pearson, 2014).

The PPVT is also supposed to screen for general verbal development and language or visual impairments (Pearson, 2014) and is thus frequently used in clinical settings (Hoffman et al., 2012). One of the key advantages of the PPVT is that it can be used with learners of very low proficiency. Also, it can be used with children aged 2 onwards as it does not require the candidate to be able to read.

Apart from considerable costs, one of the disadvantages of the PPVT lies in the ambiguity of some of the pictures that remains despite the great care taken in the test design and validation stages. Also, the item sampling and thematic grouping of items seems problematic as one or two unfamiliar areas might distort results considerably for a particular candidate. Given that there seems to be no principled rationale behind the item sampling other than a thematic one (e.g. frequency-based), it seems challenging to relate PPVT scores to other L2 vocabulary research. The PPVT thus seems of limited use in L2 vocabulary studies that seek to link findings to comprehension of written or spoken discourse, particularly coverage-based research (e.g. Adolphs & Schmitt, 2003; Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt, Jiang, & Grabe, 2011; van Zeeland & Schmitt, 2013).

More recently, a similar research tool based on the principle of testing receptive vocabulary knowledge by asking candidates to match words with a series of pictorial cues has been put forward as the Pictorial Vocabulary Size Test (PVST) (Tseng, 2013). Developed for a Taiwanese context, this diagnostic test, primarily intended for primary school children, also builds on the minimization of involvement of grammatical knowledge or reading skills by using pictures as cues for lexical multiple-choice options.

In contrast to the PPVT, candidates are presented with only one picture but four words. The target words for the PVST are sampled from a pedagogical list of 1200 high-frequency words. Based on findings from a three-parameter Item Response Theory (IRT) validation study of the test, Tseng (2013) concludes that it constitutes a highly reliable measurement tool, which, when scored with an IRT model can even account for and overcome some traditional weaknesses of the employed multiple-choice format (e.g. guessing) and thus provide a more accurate representation of vocabulary knowledge than traditionally scored MCQ vocabulary measures.

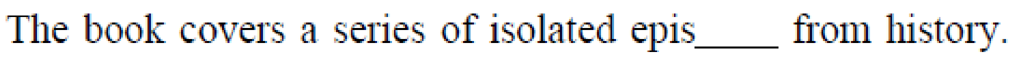
The value of Tseng's (2013) study therefore lies in highlighting the beneficial potential of IRT analyses to both validation and scoring procedures in vocabulary assessment. Tseng (2013) underlines that Classical Test Theory (CTT), despite balanced sampling from frequency bands, "does not account for the discrepancy between raw scores and item responses" (p. 71). Hence, Tseng (2013) claims that "[m]odeling item responses rather than raw scores in such [longer multiple-choice] tests not only greatly increases the likelihood of capturing the true value of vocabulary size, but also makes it possible to model the guessing phenomena of the test items" (p. 71). Being a test developed for a specific national and potentially culturally rather homogenous context, it remains to be demonstrated, however, whether the PVST and its pictorial cues could be used in other contexts with comparable success. To the best of my knowledge, no other studies employing this instrument have yet been published. Another Picture Vocabulary Size Test, developed by Nation and Anthony (2016) for use with children, is under construction at the time of writing this thesis.

2.3.1.6. Tests of productive vocabulary knowledge

The most prominent tests of vocabulary knowledge (VLT, VST, Y/N) all claim to be testing some form of receptive written vocabulary knowledge. This propensity may be the result of both principled decisions based on the interest of particularly reading researchers in the contribution of vocabulary knowledge and the perceived simplicity and practicality of designing, administering and scoring such typically selective, discrete and context-independent tests. It is important to note here, that tests that use form recall response formats, such as the CATSS, cannot automatically be classified as productive vocabulary tests.

While most measures of productive vocabulary knowledge come in the form of embedded tests or profiling free candidate production with tools such as the Lexical Frequency Profile (Laufer & Nation, 1995), some attempts have been made to measure productive vocabulary knowledge in a controlled fashion, acknowledging the usefulness of having an array of complementary tests at a researcher's disposal as different types of items tap into different areas and degrees of vocabulary knowledge (Paul, Stallman, & O'Rourke, 1990). Literature on these measures is scarce, as is their use (some notable exceptions are Laufer (1998) and Thekes (2013)). However, two examples shall be discussed briefly in the following.

One attempt to design a vocabulary size test of controlled productive ability was undertaken by Laufer and Nation (1999). Their test is based on the VLT in that it adopts a view of vocabulary frequency levels from which to sample test items and uses the exact same target words as the revised Schmitt, Schmitt and Clapham (2001) VLT versions. This Productive VLT is postulated to test controlled productive ability as a measure of vocabulary growth (Laufer & Nation, 1999). "Controlled productive ability" they thereby understand as "the ability to use a word when compelled to do so by a teacher or researcher" (Laufer & Nation, 1999, p. 37). In order to clarify the desired target for the test takers, they disambiguated alternatives by providing the minimal number of letters of the target in a cue. Resembling C-Test items, a model item would thus look as follows:



The book covers a series of isolated epis_____ from history.

Figure 7: PVLT example item (Laufer & Nation, 1999, p. 37)

It seems, however, at least debatable whether this really constitutes the minimal number of cue letters or whether the sentence context could not have been constructed in such a way as to disambiguate while still remaining non-defining. Schmitt (2010) further criticises the test in that it has not been empirically explored whether the fact that "some of the target words have only one letter to disambiguate them, while others have up to six" (p. 203) has any effect on the relative difficulty of the items.

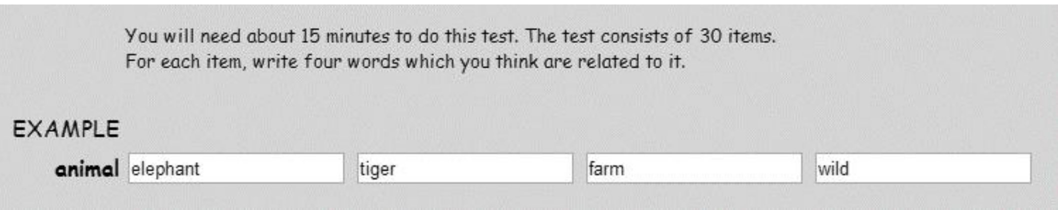
Laufer and Nation (1999) maintain that their productive VLT sufficed in terms of practicality, reliability and validity, inasmuch as it distinguished between different proficiency groups. However, this alone might not be the most solid of grounds to base a validity argument on, particularly since the scoring procedure of their study needs to be problematized.

It is not fully clear what the PVLT intends to measure. Describing the test as a form-recall measure would not be accurate as part of the form is provided for the test taker (Schmitt, 2010). Schmitt (2010) even goes so far as to claim that "the PVLT might be better considered an alternative way to measure receptive vocabulary knowledge rather than a measure of productive vocabulary" (p. 205). He further points out that the behaviour of individual items within frequency levels has not been probed and comparisons between the PVLT and the LFP are of limited meaningfulness as they are based on different frequency breakdowns (Schmitt, 2010). Laufer (1998), for instance, found no correlations at all between the PVLT and the 2,000+ level of the LFP.

In addition, not penalizing spelling mistakes in a test of productive vocabulary knowledge seems, at best, questionable, even though the authors acknowledge that productive knowledge, like its receptive counterpart, is a matter of degrees of mastery. Fitzpatrick and Clenton (2010) further criticize the amount of receptive vocabulary knowledge involved in answering the items of the PVLT through the necessary processing of the sentence context.

Given that the PVLT starts assessing at the 2,000 word frequency level only, Abdullah et al. (Abdullah, Puteh, Azizan, Hamdan, & Saude, 2013) have recently argued for a need to develop further versions of the test that capture vocabulary knowledge of lower-proficiency learners. Their 20 item PVLT500, which focuses on the 500 most frequent word families, avoided the pitfalls of inconsistent scoring and showed promising results in an initial large-scale validation study. However, low reliability estimates of this new test version remain a concern.

A different approach to assessing productive vocabulary knowledge was taken by Meara and Fitzpatrick (2000) in the development of their Lex30. Lex30 also uses word frequency as a criterion to assess vocabulary production, but does so by asking the candidate to provide four associated words for 30 target items. These results of the association task on the 30 high-frequency items, which are all taken from the first 1,000 most frequent word families, are then scored against word frequency lists (Fitzpatrick & Meara, 2004). Any answer that is infrequent, i.e. a word from any frequency band beyond the 1K level, is awarded one point, resulting in a maximum score of 120. The advantages of Lex30 are that it requires little use of receptive knowledge as only the target word is provided without any context. Care was taken in the design of the cue items, that frequency is controlled for and that they do not generate strong primary associations or association responses that are among the 1,000 most frequent English words themselves (Fitzpatrick & Clenton, 2010). A screenshot of the online test version illustrating the format with exemplary responses can be found in Figure 8.



You will need about 15 minutes to do this test. The test consists of 30 items.
For each item, write four words which you think are related to it.

EXAMPLE

animal elephant tiger farm wild

Figure 8: Lex30 example item (http://www.lognostics.co.uk/tools/Lex30/)

When validated, Lex30 yielded acceptable re-test reliability values and seemed to "be sensitive to improvements in learner's language ability at lower ranges of proficiency, but [...] not able to distinguish high-level learners from native speakers" (Fitzpatrick & Clenton, 2010, p. 544). Fitzpatrick and Clenton's findings regarding modality are also worth mentioning. While the mean scores in their study showed that candidates' performances were not significantly influenced by modality, correlation analysis revealed that the test might work differently depending on whether test takers are asked to provide the answer in written or spoken form.

It has to be mentioned, however, that Fitzpatrick and Clenton's population sample sizes were minimal for the types of analyses they conducted. Both allegedly being measures of productive vocabulary knowledge, Lex30 correlated only moderately (.5) with the PVLT (Fitzpatrick & Clenton, 2010). Walters (2012) found a higher correlation between the two measures (.77) in a replication study with candidates from a range of proficiencies in Turkey. In a different set of studies, Fitzpatrick and Clenton (2017) found no correlations between lexical frequency profiles of learner compositions or of a brainstorming task. This seems hardly convincing, despite the efforts of the authors to assert the complexity and subtle differences of the constructs underlying the different tests. Granted the fuzziness of the multifaceted construct of vocabulary knowledge, it sometimes, however, seems merely a welcome cop-out for test designers and validation researchers. Fitzpatrick and Clenton's (2010) conclusion that "'productive vocabulary knowledge' is not a precise enough construct for an investigation of validity" (p. 545), at least, appears highly questionable.

Although learners, or indeed professional linguists, might not generally succeed in estimating the frequency level of a particular word (Alderson, 2007), it seems nevertheless slightly problematic that the instructions withhold the scoring criteria from the candidates. Given the idiosyncratic nature of association task responses (Cremer, Dingshoff, de Beer, & Schoonen, 2010; Fitzpatrick & Clenton, 2010), the task to write down any associated words also begs the question not only how much control this test actually provides for the administrator or researcher but, more crucially, what exactly it is this test is attempting to assess, let alone how the scores can meaningfully be interpreted.

In terms of Nation's framework, Fitzpatrick (2007) claims that Lex30 scores provide information about a candidate's knowledge about how a word is written and spelled (written form-productive), about what word should be used to express this meaning (meaning concept-productive) and about what other words could be used instead of the targeted one (meaning associations-productive). However, even if one is willing to accept the first two construct aspect propositions, the test's instructions clearly do not ask the test taker to provide synonyms as outlined in Nation's defining question for meaning associations-productive. Even though the candidates have to *produce*, in other words *write down* their associations, the task itself is more aligned with the defining question for meaning associations-receptive: "What other words does this word make us think of?" (Fitzpatrick, 2007, p. 129). This again underlines the problematic nature of Lex30 score interpretations, which, unfortunately, renders it a tool of limited use to establish vocabulary knowledge in pedagogical and research contexts.

Stewart's (2012) multiple-choice test of active vocabulary knowledge claims to be combining the advantages of multiple-choice test (practicality in administration and scoring), while avoiding guessing effects and still measuring productive word knowledge. In his format, candidates are asked to provide the first letter of the target word after being presented with a translated definition, the second and third letter or the target word and POS information. Although the chance of successful blind guessing is reduced to 0.04% in this format, it is questionable whether it is actually a test of productive vocabulary knowledge as only the first letter of the word needs to be provided, deliberately neglecting the need for accurate spelling as part of productive word knowledge. Also, though promising, validation evidence for the test is currently scarce and not fully convincing. A validated test of productive vocabulary knowledge therefore remains a desideratum.

2.3.1.7. Tests of spoken vocabulary

The majority of vocabulary tests focus on written vocabulary knowledge, disregarding the spoken component of lexical ability (Barclay, 2013; Nation & Webb, 2011). Since the written and spoken vocabulary knowledge of learners might differ considerably, Read (2000, 2007), as well as Milton (2009) have argued for the separate assessment of vocabulary knowledge in these two modalities. It has been problematized that previous research into the relationship between vocabulary knowledge and listening had to rely on tests of written vocabulary knowledge, which might not adequately and validly represent candidates' aural vocabulary knowledge. Although Van Zeeland (2013) suggests that the gap between spoken and written vocabulary knowledge might not be as big as initially claimed based on findings by Milton and Hopkins (2006), there is grounds for developing a measurement of spoken receptive vocabulary knowledge.

Only few attempts to close this gap have been made and there is yet no fully validated test for spoken receptive (or indeed productive) vocabulary knowledge. Fountain and Nation's (2000) vocabulary-based graded dictation test has been criticized for involving too much listening skill for it to be a valid measure of vocabulary knowledge (Barclay, 2013). An aural version of the yes/no test, A_Lex, has been suggested (Milton & Hopkins, 2005, 2006; Milton, Wade, & Hopkins, 2010) but has not yet convincingly been demonstrated to yield valid scores as it suffers from the typical weaknesses of checklist tests and has more resemblance with a self-assessment tool than a test based on some sort of verification or demonstration of knowledge.

As an alternative, Barclay (2013) proposes a Vocabulary Levels Translation Test to respond to the need for an aural vocabulary test. The test has the advantage of requiring the candidate to demonstrate knowledge of the word and testing spoken vocabulary knowledge both in context and for isolated word forms. However, it might suffer from regional and validity limitations of translation formats and issues with scoring for partial knowledge. Thus, the VLTT has yet to be shown to work beyond the promising piloting phase. McLean, Kramer and Beglar's (2015) recently developed Listening Vocabulary Levels Test (LVLT) is a variation of Barclay's suggestion. The design and validation procedures, however, could be considered problematic. For example, the unusually high correspondence between the test item score and the criterion measure score from the interview could simply be due to the additional prompts in the interviews and thus the similarity of the tasks. Also, the LVLT employs the same format as the VST and so presents items in limited context.

The issue of contextualized vocabulary items might carry even more weight in aural vocabulary tests as the construct-related boundary between segmentation and lexical knowledge seems even more blurred than in providing written context. On the other hand, Read (2007) argues that an aural vocabulary test of only isolated words might be limited in its appropriateness and meaningfulness. It might also suffer from overestimation of a learner's ability to listen to continuous speech (van Zeeland, 2013). A comprehensive vocabulary test of receptive word knowledge would therefore have to account for the spoken dimension of receptive word knowledge beyond (single) isolated words and self-assessment.

2.3.2. Measurements of depth of vocabulary knowledge

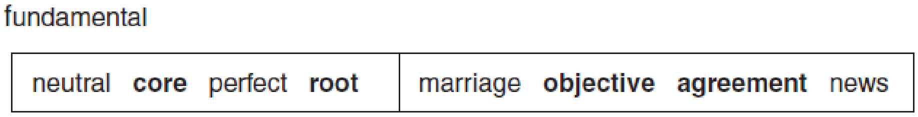
As outlined earlier in this chapter, different approaches to conceptualize and operationalize depth of word knowledge in vocabulary tests have been proposed. Some examples will be evaluated in the subsequent sections.

2.3.2.1. Word Associates Format (WAF)

In an attempt to do justice to the broadened notion of multiple dimensions of word knowledge, Read (1993, 1995) developed tests using the so-called Word Associates Format. These tests, in contrast to conventional size tests, focus on the depth or quality of word knowledge in that they measure the learner's knowledge of associations. Such associations could be of a paradigmatic, syntagmatic or analytic nature. This format could therefore complement traditional size measures of form-meaning link knowledge.

The Word Associates Format is generally presented in a variation of a multiple choice format, thus having the advantage of being economical whilst simultaneously tapping into not only meaning senses of target words but also some of a word's uses. As such, however, it also suffers from the same traditional threats to validity as conventional multiple choice tests, such as guessing effects (Read, 1998). Greidanus et al. (2004) assert that the format is a relatively efficient way of measuring deep word knowledge and that it has the advantage of being independent of the L1 of the test taker. They also maintain, however, that target selection can be fairly difficult and restricted as "[n]ot every word has the right properties to function as a stimulus word" (Greidanus et al., 2004, p. 203). This additional criterion of a word's usability as a target may introduce a confounding element, compared to tests which rely on purely frequency-based or difficulty-based sampling.

The format is based on the core notion of word association and presents the candidate with "items that consist of a target word and six or eight other words, half of which are associated with the target word and half not" (Read, 2007, p. 113). An illustrative example is shown in Figure 9.



fundamental

neutral **core** perfect **root**	marriage **objective** **agreement** news

Figure 9: WAF example item with answers in bold (Schmitt, Ng, & Garras, 2011, p. 107)

The associations, depending on the variation of the format, are primarily semantic and syntagmatic (collocational). In its original eight-option format, the options are equally distributed, also visually, across these two categories of associations and the test taker is instructed to indicate the four correct associations.

To reduce the format's susceptibility to guessing, the distribution of correct responses within each category could range from one to three acceptable options per category. In an alternative version based on the same principle, Schoonen and Verhallen (2008) suggest a six-option format which focuses on sematic relations between the options and the target words. This six-option design has been successfully adopted in several other studies (Beks, 2001; Greidanus et al., 2004; Greidanus & Nienhuis, 2001; Verhallen, Oezdemir, Yueksel, & Schoonen, 1999). Schoonen and Verhallen (2008) claim in their study that this test format is particularly suitable for researching vocabulary in young language learners as increased decontextualized semantic knowledge could be seen as an indicator for a more developed and advanced lexicon. Given this assumption, the test seemed to work well statistically in their study with 9-12 year olds. However, the distinction between correct and incorrect answers in this version where all distracters share some semantic relation to the target word appears somewhat arbitrary, as is also acknowledged by Schoonen and Verhallen (2008).

Also, there still remain issues with the test format's validity. Schmitt, Ng, and Garras (2011), comparing the test scores of participants on different WAF versions with candidate's verified word knowledge as elicited and judged in an individual interview, found that the WAF "suffers from a tendency to overestimate learner knowledge" (p. 123). Their analysis of 18 Japanese adult EFL learners showed that "interpreting split scores on the WAF is problematic" (Schmitt, Ng, & Garras, 2011, p. 109) with WAF scores paradoxically both over- and underestimating vocabulary knowledge at times. They conclude that the WAF is only a suitable measurement instrument for learners that can be located at the extreme ends of the scoring scale. This renders the instrument problematic as the majority of candidates taking a test will probably be clustering in the middle of the scoring range.

Two answering patterns were identified as particularly problematic: (1) when candidates select one correct and one incorrect associate, thus cancelling answers out, and (2) when candidates scored in the collocation section only, demonstrating no knowledge of the word meaning. This could either mean that the format is indeed highly prone to guessing or that knowledge of the form-meaning link is not necessarily a prerequisite for collocational knowledge. If the latter were the case, the validity of an assumed implicational scale between these components of word knowledge would need to be further probed. In their study, Schmitt, Ng and Garras (2011) also investigated the strategic behavior exhibited by candidates taking the WAF.

Exploring different types of distractors (Greidanus & Nienhuis, 2001; Schmitt, Ng & Garras, 2011), research has found that form-based or antonymic distractors should generally be avoided in favor of no-relationship distractors, distractors closely related in meaning to the target or associate words, particularly for more advanced learners, or distractors that can potentially pair up with one another.

Like in the Yes/No format, different scoring procedures have been debated for the WAF as well. Schmitt, Ng and Garras (2011) found in their study that "the All-or-Nothing method is probably the best for the 6-option version, and the One-Point method for the 8-option version" (p. 122) of the WAF. However, their study is limited in the generalizability of its findings due to small sample sizes of both items and participants and the relatively truncated nature of the sample.

Schoonen and Verhallen's (2008) test version also could be argued to be measuring or profiling lexicon organization rather than word knowledge. Even though Nation and Webb (2011) argue that Read's WAF format could also be classified as measuring comprehensive word knowledge as it assesses knowledge of form, meaning, concept, referents and collocation, this version in particular seems to assess primarily network knowledge (Read, 2004a).

While network knowledge certainly classifies as an important aspect of deep word knowledge, it might be more relevant to studies investigating the organizational structure of the mental lexicon than to whether or not and how well a particular word is known in terms of Nation's (2001) taxonomy. The same criticism holds true for the deep word knowledge test for advanced learners of French (Greidanus et al., 2004), which also uses the word associates format. Bogaards (2000) therefore rightly states that "[i]f its [the test format's] only purpose is to measure how well the selected target items are known, then the test may not do a very good job. But one could be interested also in more general qualitative knowledge of the lexicon" (p. 496), in which case this format should not be disqualified.

2.3.2.2. Vocabulary Knowledge Scale (VKS)

Following a developmental approach to vocabulary measurement, Paribakht and Wesche (1997) (see also Wesche & Paribakht, 1996) suggested a six-point elicitation scale ranging from "I don't remember having seen this word before" to "I can use this word in a sentence", which is complemented by a request to provide a synonym, L1 translation or sentence. This scale, which takes into account the partial and incremental process of word knowledge and combines assessing the form-meaning link as well as some aspects of depth of word knowledge, has become known as the Vocabulary Knowledge Scale (VKS). While Laufer and Goldstein (2004) argued that the VKS is an indirect test of word meaning rather than a test of vocabulary depth, it is often classified as the latter.

The VKS surpasses ordinary self-report scales in that the higher stages of the scale require not only a self-assessment but also demonstration of that self-asserted knowledge. The major problem, however, with the VKS, as with all other developmental approaches to vocabulary measurement, is that we don't know enough about the incremental acquisition of vocabulary knowledge in order to decide on the best scale to measure it (Schmitt, 2010). Even Wesche and Paribakht (1996) acknowledge the "lack of theoretical consensus about the nature and course of development of L2 vocabulary knowledge" (p. 32). Neither the number of levels or stages, nor the actual stages are grounded in a sufficient amount of empirical research to design a measurement scale which would allow for highly valid claims. Also, Schmitt (2010) argues that the scale might not be unidimensional in that it involves a "constellation of lexical knowledge" (p. 220) at the different stages, mixing receptive and productive elements in an unprincipled way and offering various degrees of contextualization. Schmitt (2010) further echoes Read's (2000) critique that the intervals between the five stages might not be equidistant. Stewart, Batty and Bovee (2012) explored the psychometric dimensionality of the VKS empirically and found a weak multidimensionality and unclear construct distinctions. The close difficulty proximity of some knowledge levels, they argue, impedes the results' interpretability and the VKS' usefulness as a diagnostic measure for educators.

In terms of demonstration of knowledge, produced sentences at the highest level, presumably showing the highest degree of mastery bear a number of scoring issues as acceptable and even sophisticated sentences could be produced by candidates that do not sufficiently demonstrate knowledge of the target word (McNeill, 1996; Schmitt, 2010). The VKS does not come with adequate scoring rubrics and guidelines that would minimize marker subjectivity at this stage (Bruton, 2009). Bruton (2009) also criticizes that it precludes L2 form recall and that in cases of homographs it is not clear for the candidate which core meaning is actually targeted.

Despite this, the VKS has frequently been used as a research tool (e.g. Bruton, 2009; de la Fuente, 2002; Horst, Cobb, & Nicolae, 2005; Paribakht & Wesche, 1997; Paribakht, 2005; Pulido, 2004; Rott, Williams, & Cameron, 2002; Wesche & Paribakht, 2009). Golonka et al. (2015) even claim that the VKS is "the most widely used scale for measuring vocabulary depth" (p. 25). Despite its operationalization of some valid assumptions, however, it may have more merit as a supplementary instrument for classroom teachers, particularly for capturing initial stages in word learning (Schmitt, 2010). Wesche and

Paribakht (1996) themselves admit that the VKS "is not suitable for testing large numbers of students in its present form" (p. 33). Schmitt and Zimmermann's simplified variation of the VKS (2002) unfortunately suffers from the same limitations in principle.

2.3.2.3. Test of English Derivatives (TED)

Schmitt and Zimmermann (2002) and Ward and Chuenjundaeng (2009) have demonstrated that derivatives are challenging for learners, even at an advanced proficiency level. The form-recall measure TED aims to tap this aspect of word knowledge. It requires candidates to provide the derivative forms of target words in sentence contexts, one per targeted part of speech. While it seems a valuable addition to the toolkit of vocabulary teachers and researchers, there is some "potential fuzziness" (Schmitt, 2010, p. 229) as to the development of the scoring key, which users need to be aware of. The TED is, however, the only available systematically designed measurement instrument available to test this aspect of word knowledge other than simple gapped grids. Nevertheless, it has not been employed extensively in vocabulary and SLA research to date.

2.3.2.4. Collocation measures

Collocation knowledge has been described as "one of the most important types of 'contextualized' word knowledge" (Schmitt, 2010, p. 229). It is therefore understandable that the field has seen an increased interest in depth measures featuring this component in recent years.

Early attempts to measure collocational knowledge employed translation formats, such as in Bahns and Eldaw's (1993) study of the English collocation knowledge of German speakers. They used German prompt sentences containing translation equivalents of 15 English verb+noun collocations and asked for their translation into English. The method, however, lacks tight control as translated sentences might be acceptable but not contain the targeted collocation. Other researchers have attempted to measure this kind of knowledge using cloze items (Farghal & Obiedat, 1995), which suffered from similar problems in that an acceptable but not the targeted collocation could be inserted by the candidate. This issue of lacking restriction becomes even more salient when the entire collocation is asked for rather than just one element of the collocation. The inclusion of initial letters in the gaps has been suggested to constrain the choices of the test takers as well as the provision of L1 translations. The latter, however, is only feasible if there is no identical collocation in the L1 so that it is really the collocational knowledge in the L2 that is being measured. Gyllstad (2007) further criticizes the sample sizes, the unprincipled sampling and lacking reliability evidence in these early studies. Schmitt (Schmitt, 1998a) used a sentence elicitation task to measure productive collocation knowledge in his research. Candidates were asked to provide three sentences per target word, each constrained by a topical context. However, he reports that his scoring criteria were probably too generous and lenient and that this format is therefore limited in its usefulness.

More recently, Bonk (2001) has investigated several collocation test formats. He also used sentence cloze items but focused on the insertion of elements of either verb+object or verb+preposition collocations. As a third measure he used a four-option multiple-choice format but in an odd-one-out design, whereby three options contained valid collocations and the candidates were asked to identify the one option which contained an incorrect collocational usage. He administered the three tests to 98 Asian EFL students and found satisfactory reliability values for all measures except the verb+preposition cloze test. The population performed similarly on all three measures and high collocation scores correlated with high scores on a test of general proficiency. Based on his IRT analysis he concludes that the verb+object cloze measure and the multiple choice format work well. The findings also suggest that collocation knowledge might be an indicator of advanced proficiency or advanced word knowledge.

In a more decontextualized approach, Mochizuki (2002) presented 54 Japanese test takers with node words and four collocation options to choose from. The findings from his study, however, render the instrument questionable as reliability values varied considerably, probably due to the homogeneity of the population. Also, the test items were not able to capture a meaningful gain in collocation knowledge after 75 hours of instruction.

Barfield (2003), adopting a developmental approach reminiscent of the VKS, designed a scale by means of which students can be asked to judge their familiarity with a particular decontextualized collocation and its frequency. In addition, his test contained non-collocations to prevent guessing. The measure yielded acceptable reliability values and discriminated well between stronger and weaker groups of test takers. Even though Gyllstad (2005) problematizes that the test contains possible non-collocations, the scale's focus on frequency should make it clear that typicality or probability rather than possibility of collocations is the appropriate criterion. Therefore, his critique does not seem justified in this respect. Much more problematic seems the scale itself, which is again a self-evaluation tool. Schmitt (2010) raises the point that the sampling of the targets lacks clarification. According to him, a rating of the highest interval is only appropriate if the target collocation is indeed highly frequent. It is, however, not clear whether all real collocations were highly frequent or how the collocations were assigned the appropriate interval if they were not (Schmitt, 2010).

Gyllstad (2005, 2007, 2009) attempted to develop two collocation test formats that focus on receptive collocation knowledge and also include the frequent category of delexical verbs. His format COLLEX 5 asks candidates to select the real verb+noun collocation from three options. The 50 test items were controlled for frequency of their components and the targets had to feature a minimum z-score of >3. A corpus analysis was used to check that the distractors were not real collocations.

Gyllstad's alternative format COLLMATCH 3, initially developed as a multiple matching grid, resembles the checklist yes/no format in its latest version. Candidates simply indicate whether a presented verb+noun combination is a collocation or not. While the initial matching format suffered from several shortcomings and was rightly discarded after the piloting, the current format of COLLMATCH seems not only misleadingly named but also prone to the same problems of traditional yes/no tests. Despite reasonable reliability values and the ability to profile clear progressions along proficiency levels of both formats, COLLEX 5 seems a much more solid and meaningful way of assessing collocation knowledge.

Eyckmans' (2009) Discriminating Collocations Test (DISCO) basically also follows the principle of a yes/no format with interspersed non-collocations. However, her longitudinal validation study is based on the results of only 25 students, which limits the meaningfulness of the investigation.

Revier (2009) criticises previous collocation knowledge assessments' (e.g. Gyllstad, 2009) reliance on "a single elicitation method that involves presenting test takers with a node-word prompt (e.g. attention) and asking them to select or supply one or more collocates (e.g. *call, draw, pay*) of that node word" (p. 125). This, he claims, only gives little or no insight into the candidate's knowledge of the whole collocation (Revier, 2009). However, knowledge of the whole collocation might be a desirable target for assessment, particularly for collocations that function as phrasemes rather than simply partners (Macis, 2013). Revier thus strongly argues for a relativisation of Nation's (2001) view of collocation as a word-property or subcomponent of word knowledge. According to Revier (2009), collocation knowledge should be viewed as independent knowledge, whereby collocations are treated as meaning units in themselves, which is why he calls for assessments that require the candidates to produce or recognize whole collocations.

He therefore suggests CONTRIX, a matrix format in which test takers construct the collocation to fill a sentence gap from potential constituents. The format provides some context and restricts responses as learners select from a limited number of choices. This provides a reasonable alternative to cloze gaps that indicate the initial letters of the targets. The 45 item test was balanced for semantic categories, verb constituency, item frequency and noun-constituent frequency. It was piloted on 56 Danish EFL learners and showed promising psychometric results. However, Revier (2009) admits that a number of individual items performed poorly and are in need of revision. Further validation evidence of an improved test appears necessary, although the format certainly bears potential. However, it seems debatable whether the test really taps productive collocation knowledge, as claimed by the author (Revier, 2009).

2.3.3. Test batteries assessing more than one aspect of word knowledge

A limited number of studies have attempted to design test batteries that measure several aspects of word knowledge. However, most of these have used a collection of (pre-existing) measures and combined them. No researcher as yet has attempted to design one integrated test battery to assess various word knowledge aspects.

Schmitt (Schmitt, 1998b) was probably the first to undertake a comprehensive investigation into the aspects of knowledge of spelling, word class, derivation, meaning(s), association and collocation. However, he measured these aspects in time-consuming interviews, which are not feasible for larger scale assessments. Also, the validity of the intervals some of his scoring scales, designed to capture the incremental nature of vocabulary learning, could be contested.

Ishii and Schmitt (2009) describe an integrated diagnostic test of vocabulary size and depth, developed for the Japanese higher education context. Their study (N=523) highlights the need for such a multidimensional diagnostic test and provides relevant insights into the relationship between different knowledge aspects as well as principled integrated scoring schemes and accessible score reporting. They devised a battery of four tests to assess vocabulary size, knowledge of polysemy, knowledge of derivative word forms and lexical choice between near synonyms. Vocabulary size was measured by means of a 75 item version of the VLT, in which the items were sampled from a lemmatized BNC frequency list for five frequency bands (2K-6K) and the options were presented in the L1. Knowledge of a word's multiple meaning senses was assessed by means of a 30 item multiple choice test where two out of five options were correct. This test focused on the 2,000 most frequent lemmas of the BNC and did not award partial credit. Derivative knowledge was measured by a simple, decontextualized grid into which candidates had to write the word forms. The test of lexical choice between near synonyms consisted of 54 gapped sentences into which candidates had to insert the word that fitted the context better. As acknowledged by the authors, this test has "the drawback that there is a 50% chance of choosing the correct answer by guessing" (Ishii & Schmitt, 2009, p. 11). Their simultaneous testing of vocabulary size and depth aspects showed that the types of knowledge are highly interrelated although the relationship is complex. However, their test battery is merely a combination or adaptation of existing tests and test formats (e.g. the VLT) without fully addressing their weaknesses. It is therefore not an attempt to incorporate measurements of the different aspects into one systematic test battery. In addition, none of the tests combined for their study has been validated in themselves, which must be noted as a severe weakness.

Webb (2005, 2007) also measured vocabulary knowledge in a multi-dimensional approach. His ten-part battery assessed learners' receptive and productive orthographic knowledge, productive knowledge of the form-meaning link, grammatical functions, collocations, associations, receptive knowledge of grammatical functions, collocations, associations, and the form-meaning link. While this battery, employing a range of different item types, certainly provides the most comprehensive insight into word knowledge to date, it suffers from severe limitations in terms of the number of items that can be targeted (10-20). It seems, therefore, that a better balance between practicality and comprehensiveness needs to be found for vocabulary tests.

2.4.Summary

Nation and Webb (2011) state that "the history of vocabulary size testing is a history of wrong turns and poor methodological decisions" (p. 220). It emerges from the analysis outlined above that this does not only hold for size tests and that there is a need in the field to develop improved or new vocabulary tests. In summary, it can thus be argued that currently existing vocabulary tests suffer from six major weaknesses: (1) focus on single words, (2) inappropriate sampling in terms of unit of counting, frequency bands and representativeness, (3) problematic or unprincipled selection of item formats, (4) favouring of written over spoken vocabulary knowledge, (5) focus on single dimensions of word knowledge, and (6) generally insufficient validity evidence.

Despite increasing acknowledgment of the formulaicity of the English language and the pervasiveness of multi-word expressions (Erman & Warren, 2000; Schmitt, 2010), vocabulary tests still neglect phraseological knowledge and focus on the assessment of single words only. These single word target items are sampled from outdated word lists and are based on problematic assumptions about the word family unit of counting. Additionally, the sampling is often frequency-based without recognition of the potentially decreasing power of frequency as a clustering factor. By the same token, the sampling within frequency bands is rarely representative. The item formats employed are chosen for opaque reasons and rarely questioned for their potential proneness to guessing or the meaningfulness of the scores they yield. Recognition formats are particularly problematic in this respect as they allow for guessing (Stewart & White, 2011) and do not resemble the kind of task a learner comes across in real life when there are no meaning or form options to choose from. Most vocabulary tests aim to measure written vocabulary knowledge and neglect the fact that spoken vocabulary knowledge is a crucial component of receptive vocabulary knowledge that might be distinct from its written counterpart. Existing vocabulary tests further focus on individual dimensions of word knowledge only, often concentrating on vocabulary size in establishing form-meaning link knowledge (Read, 2013). They thereby fail to provide an integrated, comprehensive profile of word knowledge, which would be useful for diagnostic purposes in both pedagogical and research settings. Finally, for those tests that are available, there is a dearth of validation evidence that aims to account for what any test can and cannot tell its score user. This is particularly problematic for tests that rely on validity evidence from studies with advanced language learners but are then being used with lower level learners. This dissertation aims to address some of these core issues and develop a novel, improved measurement and profiling tool for lexical knowledge.

3. Exploring the informativeness of vocabulary test item formats

3.1.Introduction

There are a number of key considerations and decisions to be made before the development of any language test. Principal among these is the purpose of the test as it determines the construct and specific aims in more detail, the item sampling and its sources, the length and form of the instrument, the scoring, score interpretation and reporting, and the test formats appropriate to elicit the desired information. Ideally, these decisions are taken after careful deliberation and are based on language testing principles, linguistic theories and SLA research findings. Nation and Webb (2011) state that "[w]hen designing a vocabulary test, careful thought needs to be given to the item type that is used to make sure that it is suited to the kind of knowledge it is supposed to measure" (p. 219).

However, in vocabulary assessment, decisions about which item format to use seem to be primarily governed by concerns for practicality rather than empirically grounded rationales. The design of even the most prominent vocabulary tests, which employ multiple choice questions, checklists or multiple matching items, respectively, appears to have been determined by what the test developers thought was feasible, without fully accounting for what any particular format and the scores it yields can and cannot tell about the lexical abilities of a test-taker.

Paul et al. (1990) state that "[t]he choice of test format depends on the type of information desired" (p. 1). In terms of a diagnostic test of receptive lexical knowledge, several types of word knowledge information (Nation, 2001) may be required from instruments that aim to be more comprehensive than those currently available. One key aspect is thereby knowledge of the form-meaning link (Laufer & Goldstein, 2004). This aspect of lexical knowledge when activated receptively in authentic reading or listening is most likely to involve meaning recall. Meaning recall formats, such as translation tasks or interviews, however, face severe practical limitations in large-scale diagnostic scenarios as the abundance and openness of possible ways of formulating the meaning of a lexical unit seems unmanageable and highly problematic in terms of rater reliability or automated scoring. Therefore, it seems necessary to investigate other test formats for their informativeness and correlation with meaning recall measures to make an informed judgment on the most useful item type(s) to be employed in the lexical knowledge measure to be designed. The study outlined in this chapter attempts to address this issue.

Very few studies have taken a direct comparative look at different item formats and their effects on vocabulary test scores for size tests. Laufer et al. (2004) investigated four different test formats, but they all tapped into different knowledge types – meaning recognition, form recognition, meaning recall and form recall. Their research focus was on these four knowledge types and their scalability in terms of strength of form-meaning knowledge, rather than the informativeness and comparability of test formats against a concurrent criterion.

Paul et al. (1990) compared the informativeness of multiple-choice, yes/no and interview formats in an L1 setting. Testing 20 high-ability and 20 low-ability readers on their knowledge of 44 multimeaning words in these three measures, they found that both the multiple-choice format (between .66 and .82) and the yes/no format (between .69 and .81) correlated significantly and highly with the interview as concurrent criterion measure. For testing breadth of knowledge, they conclude, the multiple-choice format might be the most suitable as it gives a representative indication of the knowledge students have of specific meanings of words. While they maintain that the yes/no test is also useful for testing vocabulary size, they acknowledge that "with this type of test, there is no way of ascertaining what students know about the words or *which* words they know" (Paul et al., 1990, p. 7). They do, however, also highlight the problematic influence of test taking strategies in the second part of their study, stating that "guessing" was frequently employed, particularly by lower ability students (21% of the cases). Nevertheless, this strategy was successful in only about a third of these attempted cases. Taking the two ability groups together, successful guessing made up only 5% of all the cases. They maintain that each of these formats has its advantages and disadvantages, as already outlined in

Chapter 2. They conclude that while the interview certainly can be considered "the most effective way to find out exactly what students know about specific words" (Paul et al., 1990, p. 8), it is time-consuming and difficult to administer and score reliably and thus not practical. Their findings therefore seem to suggest that the multiple-choice format might be the most suitable for testing vocabulary size and potentially also vocabulary "depth" as they conceptualize it, as it allows for control about which meaning sense is being tested. This finding is in line with Pike's (1979) earlier study on TOEFL vocabulary items, which reported that multiple-choice formats were among the most efficient of the item types investigated.

Henning (1991) conducted a large-scale study into the functioning of TOEFL vocabulary items, comparing eight different multiple-choice formats. His findings regarding the length and inference-generating quality of multiple-choice item stems as well as the embeddedness of stems and options are very insightful due to the large number of sample items and participants. Analysing the scores of 190 test takers on a total of 1040 items (80 familiarisation items and 120 items counterbalanced across eight format conditions), he found that items embedded in a reading text appeared to outperform the traditional TOEFL vocabulary item, in which the target is part of a lengthy sentence (in contrast to a complete paragraph) and needs to be matched with a synonym. Also, the results suggest that items incorporating inference-generating information and reduced mean length of stem tended to slightly outperform the traditional format in correlation analyses. However, none of the correlational differences in his study reached significance, rendering his claims relatively tenuous. Additionally, the presumption of vocabulary total scores on these experimental items as criterion measure against which to correlate format scores is questionable at best.

However, Henning's main interest lay in the effect of the degree of contextualisation the different item versions provided. The study is therefore rather an in-depth analysis of variations of one format (multiple-choice) than a comparison of several formats. Even his claims regarding the superiority of matching formats against supply formats regarding their generalization of validity are limited to the multiple-choice format alone. A direct comparison of several test formats with a focus on their correspondence to a concurrent criterion measure is therefore much needed. Only in this way, decisions about the most suitable item type for assessing form-meaning link knowledge can be empirically informed. The present study attempts to address this gap in the research and answer the following research questions:

RQ1: Do size test item formats closely match the criterion of word knowledge demonstrated in open meaning recall?

RQ2: Do any of the investigated formats also provide other useful information to score users?

3.2.Pilot study

3.2.1. Methodology

3.2.1.1. Participants

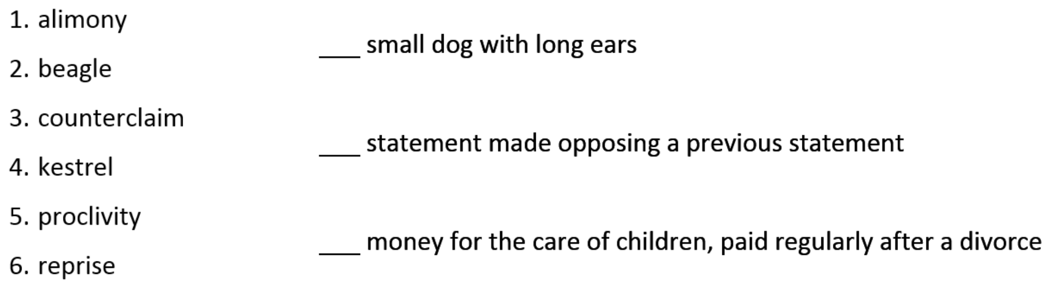
For the pilot of the research procedure, 18 English native speakers (NS) and 12 non-native English speakers (NNS), all students at a School of English at a British university served as participants. The native speakers were all undergraduate students (16 female, 4 male) with a mean age of 18.9, while 10 of the non-native speakers were postgraduate students and two undergraduate students (8 female, 4 male) with an average age of 26.9. The nine different L1s of the non-native speakers included French, Italian, Bosnian, Portuguese, Lithuanian, Greek, Arabic, Chinese and Dutch.

3.2.1.2. Target items

In preparation for a pilot study, a pretest was conducted with 15 NS, all undergraduates at a School of English at a British university, who were administered a 100 item version of Nation's VST. These 100 items were collated from the two 20K VST versions published online (Nation, 2014). Based on Nation, Goulden and Read's (1990) assertion that native speaker vocabulary size grows at about 1,000 word families per year, 50 items were taken from the section of Version A that purportedly measures knowledge of the 11K-20K range. Another 50 items were taken from the section of Version B that purportedly measured knowledge of the same frequency range. Items from the two versions were collated to increase the number of potential target items for the pilot study.

The initial study design aimed at comparing NS and NNS in the pilot. Therefore, 48 items ranging in facility value from .27 to .87 in the NS pretest were administered to 10 NNS, to ensure that there would be a mixed range of items, some of which would be known to parts of the target population, and some of which would not. From these, 36 items were selected for their average facility values, ranging from .32 to .88 across both groups with a total average facility value of .61 across all items and groups. This way it was hoped to have a spread of results regarding the pilot population's knowledge of individual items. Also, due to the clustering of items into groups of three in the VLT multiple matching format, part-of speech was considered in the target selection, resulting in eight noun clusters, one verb cluster and three adjective clusters. As much as possible, frequency level according to a more up-to-date BNC-COCA frequency list (Nation, 2004) was factored in in the decision, making sure that items would not span across more than four 1K frequency bands on average (e.g. clustering items together that were from the 14, 15, 16 and 17K). This, however, proved rather challenging due to the items' pretest facility values, so that in two cases items that differed up to seven 1K frequency bands had to be clustered together, indicating that frequency might be a negligible factor influencing item difficulty at this low end of the continuum anyway.

Four item types were developed for each of these 36 target words: one targeting form recognition, one targeting meaning recognition and two targeting form recall. One multiple matching (MM) format in the form of the VLT and one four-option multiple choice (MC) format in the form of the VST were chosen as the recognition format. These formats were selected for the purpose of this study as they are frequently used in a range of vocabulary tests, most notably in what are perhaps the three most prominent vocabulary tests VLT, VST and CATSS.



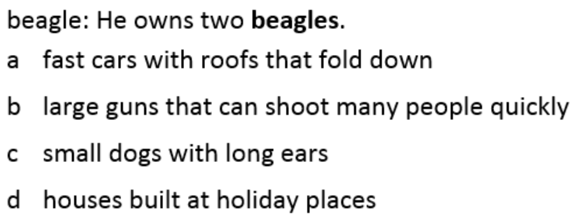
1. alimony
2. beagle
3. counterclaim
4. kestrel
5. proclivity
6. reprise

___ small dog with long ears

___ statement made opposing a previous statement

___ money for the care of children, paid regularly after a divorce

Figure 10: Example form recognition item in VLT (multiple matching) format



beagle: He owns two **beagles**.

a fast cars with roofs that fold down

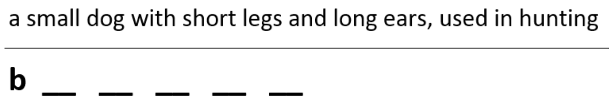
b large guns that can shoot many people quickly

c small dogs with long ears

d houses built at holiday places

Figure 11: Example meaning recognition item in VST (multiple-choice) format

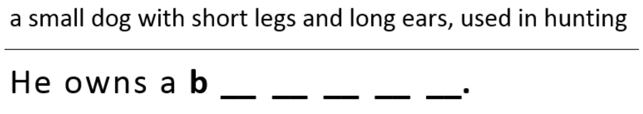
Two form recall formats providing a definition of the target word as well as the initial letter and an indication of the number of letters of the target word to disambiguate were used as recall formats. One of these form recall types gave the definition of the target word, its first letter and the blanks only (DEF). The second form recall type additionally presented the target word in a non-defining short sentence context (CON). An alternative recall format with no indication of the length of the target word, as used in the CATSS test (Laufer et al., 2004), was considered but discarded before the pilot as it showed a highly problematic proneness to ambiguity in candidate answers.



a small dog with short legs and long ears, used in hunting

b __ __ __ __ __

Figure 12: Example form recall format (definition only)



a small dog with short legs and long ears, used in hunting

He owns a **b** __ __ __ __ __.

Figure 13: Example form recall format (with context sentence)

The 36 target items were clustered into four groups of 9 items, which were balanced according to pretest facility values. Four test versions were then drawn up, which featured all four item clusters in all item type modalities in a Latin Square design. For example, the target word "beagle" was presented as a MC item in Version A, as a MM item in Version B, as a recall item without context in Version C and as a recall item with context in test Version D (see Appendix A for all test items).

Table 1: Item cluster distribution across formats and test versions

	Version A	**Version B**	**Version C**	**Version D**
Multiple Matching (MM)	1-9	28-36	19-27	10-18
Multiple Choice (MC)	10-18	1-9	28-36	19-27
Form recall with definition only (DEF)	19-27	10-18	1-9	28-36
Form recall with definition and context (CON)	28-36	19-27	10-18	1-9

3.2.1.3. Procedure

The four test versions were administered individually to the participants as paper-and-pencil versions. After each candidate had taken the test, they were interviewed face-to-face by the researcher, probing their word knowledge on three dimensions. Candidates were asked to produce the correct pronunciation of the target word, recall the precise meaning of a word and produce at least one sentence with a typical collocation or sentence context. In doing so, it was hoped to explore whether a correct score on one of these item formats could be taken as full representation of the multidimensional nature of word knowledge or whether different items testing these aspects would indeed be necessary in a more comprehensive vocabulary knowledge measure. Also, it aimed to probe which item type could provide the most informative picture of word knowledge and best represent the meaning recall knowledge of candidates.

Interviews were chosen as the criterion measure. Although they are time-consuming, they "have the value of being a stringent unguided test of knowledge" (Nation & Webb, 2011, p. 216). However, it has to be acknowledged that the criterion measure could be argued to be somewhat different in construct than the receptive written knowledge mainly targeted in the paper-pencil test, in terms of both modality and depth. Still, the interview seemed to provide the best option to verify the meaning recall knowledge and gather some additional information on crucial word knowledge aspects. Also, when processing written vocabulary receptively in reading, learners might access different levels of processing. While partial knowledge (e.g. knowing that a beagle is a type of dog) might be sufficient in some reading contexts, more precise knowledge (e.g. how a beagle looks) may be required in other contexts. The level of knowledge required surely depends on the reading purpose in a case by case basis. However, because such a relativity assumption seems unfeasible to operationalize, it was decided to opt for precise knowledge in the meaning recall measure. While this might appear to neglect the incremental nature of vocabulary acquisition to some extent, it helps draw a much clearer and precise picture of a candidate's word knowledge. It also facilitates the interpretation of scores as this level of knowledge would allow fluent reading.

3.2.2. Results

As illustrated in the figure below, cases were labelled as "match" if the candidate was either awarded a point in both of the vocabulary test item and the respective meaning recall measure (A) or if the candidate answered neither correctly (D). If candidates were awarded a point in the vocabulary test item but did not show sufficient knowledge of the item in the open meaning recall measure, this was a case of overestimation (B), i.e. the test score overestimating the actual word knowledge of a candidate. Vice versa, if a candidate was not awarded a point in the test item but was judged to actually know the meaning of the word in the meaning recall, the test item seemed to underestimate the word knowledge (C).

		Meaning recall measure	
		known	not known
Test item	correct	Match (A)	Overestimation (B)
	incorrect	Underestimation (C)	Match (D)

Figure 14: Contingency table of matching/mismatching results

Both "overestimation" and "underestimation" are errors in measurement in that there is a mismatch between the test score and a candidate's "verified knowledge" (in this case represented by the score in the criterion meaning recall measure). While both cases can be subsumed as mismatching cases and thus signify a problem with the measurement tool, they do represent two very different item behaviours that warrant closer analysis, which will be undertaken in the main study.

The overall results did not differ markedly between the native speaker and the non-native speaker group. This might have been due to the relatively high proficiency level of the non-native speakers. Judging from the meaning recall scores, individual target items performed very differently. However this did not have an impact on the overall results when comparing the meaning recall measure with the responses in the varying item formats. For instance, while no native speaker could recall the meaning "didactic", 83% of the non-native speakers knew the word. Vice versa, all native speakers knew the word "scrunch", but only 42% of the non-native speakers could recall the meaning of that word. However, the overall percentages of matching and non-matching cases was very similar across the language groups in the varying formats as can be seen in Table 2 below.

Table 2: Comparison of pilot results for NS and NNS groups

	Recognition MM	**Recognition MC**	**Recall Definition**	**Recall Context**
Match NS only	77.8%	85.2%	63.0%	69.1%
Match NNS only	85.2%	79.6%	66.7%	72.2%
Match total	80.7%	83.0%	64.4%	70.4%

For this reason, results were collated and are presented as one group of participants in the following.

Table 3: Correspondence between test formats and criterion measure for form-meaning link

	Recognition MM	**Recognition MC**	**Recall Definition**	**Recall Context**
Match	80.7%	83.0%	64.4%	70.4%
No match	19.3%	17.0%	35.6%	29.6%

For knowledge of the form-meaning link, the results indicated that for these low-frequency targets, the MC best represents the word knowledge of the candidates. The scores of the paper-and-pencil test matched the meaning recall scores in 83% of the 270 cases (36 words split into 4 clusters of 9 targets tested on 30 participants, i.e. between 6 and 8 candidates each).

Table 4: Detailed analysis of matching/non-matching cases for criterion measure and form-meaning link items

	Recognition MM	Recognition MC	Recall Definition	Recall Context
Match (point)	53.0%	48.1%	15.6%	21.5%
Overestimation	18.1%	11.5%	0.7%	0.7%
Underestimation	1.1%	5.6%	34.8%	28.9%
Match (no point)	27.8%	34.8%	48.9%	48.9%

While the recognition measures seem to have a problem with overestimation (18% of the cases in MM format and 12% of the cases in MC format), the recall measures appear to have a particular problem with underestimating the word meaning knowledge of candidates (35% vs 29%, respectively). Although this tendency could be expected, the scope of it is somewhat surprising. The advantage of the recall measures of preventing guessing, does not seem to come into effect to the extent desired. Rather, it underrepresents the meaning knowledge of candidates for words of such a low frequency. It appears that the very nature of these words makes form recall measures even more challenging and the gap in strength of knowledge between meaning recall and form recall even larger than found by Laufer et al. (2004).

The findings also showed that, despite pretesting and careful selection, the target words might have been too challenging for the sample population. Even in the recognition formats, candidates did not arrive at the correct answer in either the test item or the criterion measure in 28%, and 35% respectively, of the cases. In the recall measures, being the more challenging ones, this value reaches 49% in both formats. The results might be very different with higher-frequency targets.

In terms of the candidates' ability to pronounce the target words correctly, the findings in Table 5 also show that the recall measures' scores do not represent this aspect of word knowledge well.

Table 5: Correspondence between test formats and criterion measure for pronunciation

	Recognition MM	**Recognition MC**	**Recall Definition**	**Recall Context**
Match	73.7%	65.2%	31.9%	35.9%
No match	26.3%	34.8%	68.1%	64.1%

This, however, may be due to the relative difficulty of the recall measures and the relative ease with which, particularly native speakers, managed to pronounce words correctly even when they did not know their meaning (see Table 6).

Table 6: Detailed analysis of matching/non-matching cases for pronunciation and form-meaning link items

	Recognition MM	**Recognition MC**	**Recall Definition**	**Recall Context**
Match (point)	65.9%	54.8%	16.3%	21.9%
Item>Pronunciation	5.2%	4.8%	0.0%	0.4%
Item<Pronunciation	21.1%	30.0%	68.1%	63.7%
Match (no point)	7.8%	10.4%	15.6%	14.1%

In general, all formats "underestimated" the candidates' ability to produce the correct pronunciation of a word to some extent, the recall measures doing so in greater magnitude than the recognition formats. In 68% and 64% of the cases, respectively, the candidates could pronounce the word correctly but did not score the point in the form recall test items. In the recognition items, this Item<Pronunciation mismatch was considerably lower at 21% and 30%, respectively. While this finding might point to a word's pronunciation being

one of the first and easiest aspects of word knowledge to be acquired that coincides with a weaker degree of form-meaning knowledge as tapped into by the recognition formats, it does not seem to allow for any meaningful inferences about a person's ability to pronounce a word based on their form-meaning link test score in any format.

Further, the pilot showed that the difficulty of the recall measures at this particular low-frequency end of the spectrum again makes it problematic to interpret a form-meaning link test score to also imply collocational knowledge of a word. This is illustrated in Table 7.

Table 7: Correspondence between test formats and criterion measure for collocation knowledge

	Recognition MM	**Recognition MC**	**Recall Definition**	**Recall Context**
Match	80.4%	82.6%	63.0%	67.0%
No match	19.6%	17.4%	37.0%	33.0%

Generally, collocational knowledge seems better represented by the individual item formats than pronunciation knowledge. However, again there is an unsatisfactory mismatch between the scores on the informal collocation measure and the form recall formats (see Table 8).

Table 8: Detailed analysis of matching/non-matching cases for collocation and form-meaning link items

	Recognition MM	**Recognition MC**	**Recall Definition**	**Recall Context**
Match (point)	55.2%	53.0%	15.2%	22.2%
Item>Collocation	15.9%	6.7%	1.1%	0.0%
Item<Collocation	3.7%	10.7%	35.9%	33.0%
Match (no point)	25.2%	29.6%	47.8%	44.8%

The fact that three of the formats (MC, DEF, CON) "underrepresented" the collocational knowledge and in the recall formats severely so, again indicates that there is little to be inferred about a person's collocation knowledge from their form-meaning link test scores. In 11% (MC), 36% (DEF) and 33% (CON) of cases respectively, candidates were not awarded the point in the test item but showed collocational knowledge in the interview. Interestingly, the multiple matching format had the reverse problem. In 16% of the cases, candidates answered the multiple matching item correctly but demonstrated insufficient collocation knowledge in the interview. Overall, however, the mismatch between item scores and collocation score, regardless of the format, is rather high, which would warrant a separate measure of collocation knowledge in a vocabulary test as the form-meaning link measure alone does not seem to yield much useful information about this word knowledge type.

For two of the three criterion measures, the MC format emerged as the one best representing or rather implying these word knowledge aspects through its scores. The pilot study, however, revealed the following issues that potentially confound implications:

- The target words selected proved to be fairly difficult, particularly in the recall measures. Frequency may therefore not be a strong clustering factor at this low-frequency end.
- Derivative knowledge could not be successfully integrated into the interview measure as the pretested low-frequency words simply did not have enough derivative word family members to incorporate this word knowledge aspect meaningfully.
- Many of these low-frequency words do not have highly typical collocates. The fact that these occur generally rarely in even large corpora means that a consistent collocation patterning is even more difficult to establish for many of these items.

It was therefore decided that some adjustments should be made for the main study based on the findings and issues in the pilot. This will be outlined in the next section.

3.3.Main study

3.3.1. Methodology

3.3.1.1. Participants

I decided to use intermediate EFL learners as a target population in the main study since these EFL learners would be the primary intended test users of the resulting diagnostic instrument. In a first round of data gathering, 80 Austrian EFL learners in their penultimate year of secondary education, who had not taken part in any of the pretests, were identified as study participants. To increase the confidence in the results with a larger population, a second wave of data gathering was conducted 3 months later with 19 Austrian EFL students starting their final year of secondary education after the summer break, thus a very similar population to the sample of the first round. Only 10 participants indicated an L1 different to German. The mean age of the participants was 16.9 years, 56 were female, 31 male (12 did not indicate).

3.3.1.2. Target items

In order to select appropriate target words at the language level of the main study participants, a reduced version of the VLT was administered to 25 Austrian EFL learners in their penultimate year of secondary education. This version consisted of the 90 items of the 2K, 3K and 5K levels of Version A of the VLT (Schmitt et al., 2001). When interpreting the findings, the experience from the pilot that facility values might drop considerably when moving from this recognition format to recall formats, was taken into account. Reordering the targets according to their BNC-COCA frequency level, it was therefore concluded that the 3K level might be the most appropriate frequency level for target sampling.

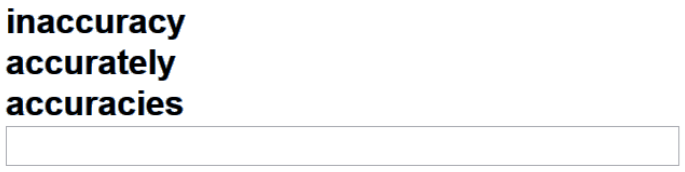
Table 9: Pretest mean facility values (N=25)

	1,000	**2,000**	**3,000**	**4,000**	**5,000**
	(k=11)	**(k=22)**	**(k=27)**	**(k=16)**	**(k=7)***
Mean FV	.82	.85	.72	.66	.49

*7 items of the revised VLT are above the first 5K according to the BNC-COCA lists

Taking the insights from the pilot and the various pretests into account, 36 items were selected for the main study and test items in four different item types developed for each of them. The targets were selected for their part-of-speech as well as their derivational forms. Four test versions were then drawn up, again featuring all items in all item type modalities between them in a Latin Square design identical to the pilot study described above. Whenever definitions or context sentences were involved in a format, this was checked against the BNC-COCA lists to ensure that these were of a higher-frequency than the target words.

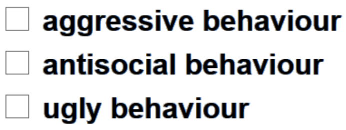
In addition to the form-meaning test versions, a test of derivative knowledge was designed testing receptive derivative knowledge of these 36 target items. This, it was thought, would represent the demands put on readers and listeners more closely than existing tests of derivative knowledge, which ask the test taker to produce the appropriate derivative form. If a reader comes across a word family member, they have to recognize or establish that this derivative form is related to a particular headword. For each of the target words, three derivational forms (sampled from the BNC-COCA frequency lists) were therefore given to test takers, asking them to write down the headword on which these were based. It was decided that three derivational forms (rather than two or merely one) would aid in disambiguating the target form. A test of derivative knowledge seemed particularly necessary to probe the notion of word families as counting units for vocabulary tests further.



inaccuracy
accurately
accuracies

Figure 15: Example item to test receptive derivative knowledge

Also, it was decided to replace the oral and rather informal collocation measure of the interview with a more formally stringent collocation format. Since only word partnerships, in which the component parts, and particularly the node as the target word, retain their literal meaning, were of interest in this present study, rather than phrasemic collocations that take on meaning(s) of their own as a larger unit, it was decided to use a format pertaining to testing form knowledge rather than meaning knowledge. Eyckmans' (2009) discriminating collocations (DISCO) format thereby seemed a useful way to operationalize the construct of interest. In this format, three collocations are presented to the test taker, one of which, however, is a non-collocation. Subjects are then asked to select the two natural and frequently occurring collocations. One test item for each of the 36 target words was created in this format. The two acceptable collocation options had to feature a minimum MI score in the COCA corpus (Davies, 2008-) of 3 and a minimum frequency count of 10 in the same corpus. Selection of the two correct options was then guided by part-of-speech as the options had to be all from the same word class (see example below). This means that it was not necessarily the two top collocations in terms of MI score and/or total collocation frequency that were selected as correct options, but a balance had to be struck in each case individually. Also, care was taken to select component parts that were below the targeted frequency band of the node word, so that sometimes the top collocations had to be discarded for that reason. All collocations, however, met the minimum criteria outlined above. The non-collocations, while semantically plausible, were checked against the COCA data to make sure they did not occur as a partnership or did not occur more than once in the entire corpus. The component parts of the non-collocations were also controlled for their individual frequency level.



□ **aggressive behaviour**
□ **antisocial behaviour**
□ **ugly behaviour**

Figure 16: Example DISCO item to test collocational knowledge

3.3.1.3. Procedure

To streamline the research procedure and increase the number of participants, the interview as a criterion measure for meaning recall was replaced by a written meaning recall measure adapted from Zhang (2013). A separate pretest using 21 Austrian EFL learners from the eventual target population revealed that the two measures are reasonably similar to justify this modification for practical purposes. In this pretest, candidates were given a list of 40 randomly selected words from the BNC-COCA 3K level, as indicated by the abovementioned trial. The instruction given was to describe the meaning of each word as precisely as possible. An example of precise knowledge description was given orally by the researcher. Afterwards, each participant was interviewed on the meaning recall knowledge of these words in an individual face-to-face session. Interviews were again recorded to enable checking the researcher's scoring. Minimally required meaning knowledge was determined by the researcher in advance. Scoring judgments were based on this. The answers in both measures (written and oral) were judged as precise meaning knowledge, partial meaning knowledge or no knowledge of the word's meaning. For instance, for the target word "bench", the ideas of an object for seating and that this object seats more than one person because it is slightly longer or bigger than a chair was defined as the minimum required for a candidate to be awarded the full point for precise meaning knowledge. Candidates who mentioned the idea of an object for sitting down but did not mention anything about the object's size, were credited with partial knowledge. For polysemous words, any one meaning was accepted as long as it was precise enough.

Table 10: Pretest results – match between written and oral meaning recall criterion measure

		ORAL		
		precise	partial	no knowledge
WRITTEN	precise	55.95%	2.26%	0.36%.
	partial	11.07%	9.64%	0.83%
	no knowledge	1.90%	1.31%	16.67%

In total, out of 840 answers (40*21) given, candidates' scores matched in 82.3% of the cases. Ruling out partial knowledge and reducing it to a 0-1 dichotomy, there is a match in the two measures in 84.4% of the cases. Correspondence reaches between 86 and 88% if "outlying" individual items or participants are removed. The advantage of the probing of the interview showed in 11% of the cases. In these, students were only given partial knowledge credit in the written measure, but demonstrated sufficient word meaning knowledge to be awarded a full point in the interview. The interview was expected to outperform the written measure in this respect as it allows for additional clarification questions of the researcher. However, this amount of underestimation in the written measure was judged to be within reason and the two measures were judged to yield similar enough scores to justify using the more efficient written meaning recall measure in future studies, especially as it would enable substantially higher numbers of study participants.

All measures were incorporated into a web-based survey tool and administered online, creating four test versions. The procedure was piloted with 17 Austrian EFL learners from the target population. Minimal changes were made in the instructions and in one example, mainly to encourage test takers to provide the fullest possible demonstration of word knowledge in the written meaning recall measure. Also, one target word was replaced as it seemed too challenging for the population, even in the recognition formats (see Appendix B for all items). Students were therefore presented with the following test variations in the main study data gathering:

Table 11: Order of tests in main study

	VERSION A	**VERSION B**	**VERSION C**	**VERSION D**
1	**Test of receptive derivative knowledge (DER)**			
2	**Form-meaning A**	**Form-meaning B**	**Form-meaning C**	**Form-meaning D**
3	**Written Meaning recall measure (MR)**			
4	**Collocations test (COL)**			

All participants started the test battery by answering the 36 items of the test of receptive derivative knowledge (1). Then, in a randomized fashion, each participant took one of the 4 form-meaning link knowledge tests (2), each of these containing all target items but in different formats as outlined in the pilot study section above. All participants then completed the same written meaning recall measure (3), before finishing the 36-item collocations test (4), which was again identical for all participants. The participants were administered all measures in a one-hour session for practical reasons. Although any cross-contamination between these steps in the battery cannot be ruled out with certainty, the order, design and amount of the items was chosen to limit interference between the individual tests. On average, participants took 38 minutes to complete all tests. Answers were afterwards coded 0 for incorrect answers and 1 for correct answers, with no partial credit given.

3.3.2. Results

The seven measures (Derivations, Collocations, Written Meaning Recall and the 4 form-meaning link test versions) performed well in terms of their reliability. The Cronbach alpha values can be seen below in Table 12.

Table 12: Reliability indices of instruments in main study

	DER	MR	COL	Form-Meaning A	Form-Meaning B	Form-Meaning C	Form-Meaning D
Cronbach's alpha	.84	.94	.96	.91	.93	.92	.92

The collocations test featured the highest Cronbach alpha value at .96, the alpha of the meaning recall measure was .94 and even the derivational knowledge measure yielded a very satisfactory alpha of .84. The individual test versions' reliabilities, featuring 9 items in each format (=36 items in total), were between .91 and .93. Reliability within the formats ranged from an average alpha of .65 (MM) to .68 (MC), to .80 (CON), to .84 (DEF) as can be seen in the table below. This is satisfactory given the lower number of items relative to the derivative, collocation and meaning recall measures.

Table 13: Reliability indices by item format

	MM	MC	DEF	CON
Cronbach's alpha in Version A (k=9)	.45	.76	.86	.81
Cronbach's alpha in Version B (k=9)	.81	.73	.88	.86
Cronbach's alpha in Version C (k=9)	.60	.78	.82	.75
Cronbach's alpha in Version D (k=9)	.75	.45	.78	.76
Average	.65	.68	.84	.80

The facility values of the instruments also confirmed that the measures were appropriate in terms of difficulty level as they seemed manageable, yet successfully providing a spread of mixed results. The DER subtest showed an average facility value of .64, being slightly easier than the meaning recall test (.58) and the collocations test (.41). Of the different item formats, the recognition formats were found to be easier for the candidates with average facility values of .77 in both the MM and the MC format. As expected, the values for the recall formats were lower at .52 (CON) and .45 (DEF), respectively.

Table 14: Average facility values

	DER	MC	COL	MM	MC	DEF	CON
Av. facility values	.64	.58	.41	.77	.77	.45	.52
(SDs)	(.24)	(.16)	(.15)	(.18)	(.18)	(.21)	(.23)

As a first step, student responses on the different formats were compared with their scores on the concurrent criterion measure of the written meaning recall. The aim of this was to find out which item format best represented the verified meaning recall knowledge. The results of matching and non-matching cases (N=891, 99 candidates x 9 items per format) is shown in Table 15.

Table 15: Match/mismatch between item formats and criterion meaning recall measure

	MM	MC	DEF	CON
Match	74.5%	77.3%	74.4%	74.3%
Mismatch	25.5%	22.7%	25.6%	25.7%

The table shows that none of the item formats functions very well in estimating the breadth of vocabulary knowledge in terms of true measurement. Almost all formats show a mismatch between test item score and criterion measure score in around 25% of the cases, with the MC format performing slightly better than the other formats at a matching rate of 77.3 percent.

Contrary to the pilot study results, the discrepancy in percentages of matching and non-matching cases between the formats is negligible. The item types all seem to perform fairly similarly on a general level, i.e. in terms of matching the criterion measure. However, when looking more closely at the results, one can see that the formats behave very differently even though the overall percentages are almost identical on a surface level.

Table 16: Analysis of matching/mismatching cases between items formats and criterion meaning recall measure

	MM	MC	DEF	CON
Match (point)	54.9%	56.7%	38.5%	42.2%
Overestimation	22.2%	20.3%	6.5%	10.0%
Underestimation	3.3%	2.4%	19.1%	15.7%
Match (no point)	19.6%	20.7%	35.9%	32.1%

Unsurprisingly, the recognition formats generally overestimate learner's word knowledge, while the recall formats tend to underestimate the "verified" word knowledge of test takers. In the MM format, overestimation occurred in 22.2% of the cases, while the MC format overestimated the candidates' word knowledge in 20.3% of the cases. Vice versa, the definition form recall format underestimated their word knowledge in about the same number of cases (19.1%). The form recall format with context performs somewhat unpredictably with almost as many cases overestimating verified word knowledge (10%) as underestimating it (15.7%).

In a further step the test scores were also compared to the test takers' answers given in the derivations test to probe whether the different test item formats yielded any valuable information about other aspects of word knowledge. The contingency table of results can be found below in Table 17.

Table 17: Match/mismatch between item formats and derivation test

	MM	MC	DEF	CON
Match	63.6%	64.0%	62.7%	67.3%
Mismatch	36.4%	36.0%	37.3%	32.7%

Similarly to the meaning recall, no item format succeeds in fully representing the derivational knowledge of the candidates. Arguably, these form-meaning link knowledge items do not necessarily target derivational knowledge. However, the notion of word families functioning as the basis of such tests would somehow imply this knowledge aspect to be captured to some extent. While the form recall format with context comes out as the format best representing derivational knowledge with 67.3% of 891 matching cases, the other formats yielded very similar overall results with 64% (MC), and 63.6% (MM) and 62.7% (DEF) matching cases. Again, at a closer look it emerges that the formats behave in a very individual fashion as regards to their relationship with this type of word knowledge. The picture is similar to the meaning recall comparison, albeit somewhat less clear.

Table 18: Analysis of match/mismatch between item formats and derivation test

	MM	MC	DEF	CON
Match (point)	52.2%	52.6%	35.5%	41.8%
Item>derivation	24.9%	24.4%	9.5%	10.4%
Item<derivation	11.4%	11.7%	27.7%	22.2%
Match (no point)	11.4%	11.3%	27.3%	25.6%

When comparing the derivational measure with the meaning recall measure (3564 cases = 99 candidates x 36 items), it was found that there is no clear inference about a person's knowledge of derivative forms that can be drawn from their "verified" form-meaning link knowledge. Table 19 illustrates this.

Table 19: Match/mismatch between form-meaning knowledge and derivational knowledge

		Form-meaning knowledge	
		known	not known
Derivational knowledge	known	43%	21%
	not known	16%	21%

While in 64% of the cases candidates knew either both the form-meaning link and the derivational forms (43%) or neither (21%), there is a mismatch in 37% of the cases. In 21% of the cases candidates could form the base word of the derivational variations without demonstrating knowledge of the meaning of that base word. On the other hand, candidates sometimes knew the meaning of a word but could not connect the derivational forms to that base word in the derivation test (16%). This could indicate that there is not enough grounds to make substantial inferences about the derivational knowledge of a candidate from their form-meaning link knowledge as demonstrated in any of the investigated item types. If information about a person's derivational knowledge is required, it probably needs to be tested in a separate derivation test item format.

In a similar vein, candidates' collocation test scores were compared with the individual item formats. The representation of this knowledge aspect through a form-meaning link item type is even weaker with scores matching in a maximum of 61.4% of the cases (DEF) and in only 59.1% (CON), 53.5% (MC) and 51.7% (MM) of the cases respectively.

Table 20: Match/mismatch between item formats and collocations test

	MM	MC	DEF	CON
Match	51.7%	53.5%	61.4%	59.1%
Mismatch	48.3%	46.5%	38.6%	40.9%

Again, the close analysis of matching and mismatching cases reveals the recognition format's problems with overestimation (41.9% and 40.7%) and the "unpredictability" of the mismatches in the recall formats.

Table 21: Analysis of match/mismatch between item formats and collocations test

	MM	MC	DEF	CON
Match (point)	35.2%	36.3%	22.8%	25.7%
Item>Collocation	41.9%	40.7%	21.7%	26.5%
Item<Collocation	6.4%	5.7%	16.8%	14.4%
Match (no point)	16.5%	17.3%	38.6%	33.4%

The implications of these findings will be discussed in the following section. Like the derivational knowledge, collocational knowledge does not seem to be easily inferable from a candidate's scores on a form-meaning link test. This suggests that a separate collocation test would be required, should information about this knowledge type be desired.

Because derivational knowledge and collocational knowledge are both types of vocabulary "depth", their relationship was further probed. In a first step, answers to the derivation measure were compared to those given in the collocation measure.

Table 22: Match/mismatch between derivation and collocation scores

		Collocations	
		known	not known
Derivations	known	28%	36%
	not known	13%	24%

The relationship between these two types of word knowledge, however, seems to be even less stable than the relationship between derivational knowledge and form-meaning knowledge. It could be argued that this is perhaps the case because the latter two share more knowledge features than derivation and collocation knowledge, in that both derivation and form-meaning link knowledge focus on the form of words to some extent, while collocations are very much associated with vocabulary use.

On a more general level, the three types of knowledge tested in this study were then compared. Taking the meaning recall measure as the "best" measure of form-meaning link knowledge, the results of this measure were related to the scores in the other two measures to probe whether there is some kind of implicational hierarchy between the word knowledge aspects. In only 23% of the cases, candidates answered all three word knowledge aspect test items pertaining to one target correctly. It was found, however, that in 75% of the cases in which the COLL measure was answered correctly, form-meaning link knowledge was also demonstrated. In 69% of the cases in which the COLL measure was answered correctly, the person also scored on the respective DER test items. Of the cases that answered the MR items correctly, 73% also answered the respective DER items correctly, while only 53% also scored on the respective COLL measure items. 67% of the people who answered a DER item correctly also answered the respective MR item correctly. The proportion drops to 44% of the cases that scored on the DER items and also the COLL items, as can be seen in Table 23.

Table 23: Analysis of correct answers - proportions of other aspects known

	Collocations known		Meaning known		Derivatives known
Proportion of cases that also knew derivatives	69%	Proportion of cases that also knew derivatives	73%	Proportion of cases that also knew meaning	67%
Proportion of cases that also knew meaning	75%	Proportion of cases that also knew collocations	53%	Proportion of cases that also knew collocations	44%

If a hierarchy of learning or mastery was assumed between the three tested knowledge aspects, this should show in the score patterns of candidates. Hypothetically, if the progression was derivative knowledge before meaning knowledge and then collocation knowledge, scores should generally follow one of the four patterns outlined below.

Table 24: Acceptable score patterns in a model DER>MR>COL

	DER	MR	COL
Pattern A	1	1	1
Pattern B	1	1	0
Pattern C	1	0	0
Pattern D	0	0	0

Scores in the following pattern would argue against this hypothetical hierarchy:

Table 25: Potential score patterns violating the assumed model DER>MR>COL

	DER	MR	COL
Pattern E	0	0	1
Pattern F	0	1	1
Pattern G	0	1	0
Pattern H	1	0	1

At look at the data in Table 26 reveals that most cases follow the pattern DER>MR>COL. 74% of the 3,564 cases are in one of the patterns A-D. Only 66% of the cases would follow acceptable patterns in a MR>DER>COL model.

Table 26: Results of score pattern analysis

	DER	MR	COL	Cases	%
Pattern A	1	1	1	828	23
Pattern B	1	1	0	691	19
Pattern C	1	0	0	578	16
Pattern D	0	0	0	555	16
Total				2652	74
Pattern E	0	0	1	182	5
Pattern F	0	1	1	268	8
Pattern G	0	1	0	286	8
Pattern H	1	0	1	176	5
Total				3564	100

The implications this might have for vocabulary development theories and test development will be discussed in the next section.

3.4. Discussion

None of the tested formats managed to demonstrate sufficiently well that their use in vocabulary size measures is unquestionably justified. Comparing the results of the pilot study with the findings of the main study, one might suggest that it was indeed the very low frequency of the target words, which was problematic and partly caused the divergence between the recall and the recognition formats. Harking back to Laufer et al.'s (2004) findings, it seems that the gap in terms of the strength of form-meaning link knowledge opens up particularly at the lower end of the frequency spectrum. This intuitively makes sense as low frequency words might be ones that are very rarely receptively encountered, and even more rarely used productively by participants. This echoes Schmitt's (2014) assertion that "as the frequency level decreases, the recognition-recall gap increases" (p. 924). It could also be taken to suggest that frequency might be a fairly random, unpredictable and weak clustering factor at this end of the spectrum.

What is similar between the pilot and the main study findings is, however, that all formats seem to feature an error in measurement of at least 20-25%. In the more robust findings of the main study, all formats misrepresented the verified word knowledge of participants in about 25% of the cases. This is highly problematic from a testing point of view. As it seems unlikely that other formats would yield better results, it calls into question whether this amount of error in measurement might just need to be accepted but adjusted for in size estimates. Particularly in the recognition formats the overestimation appears to be fairly systematic, which at least offers the potential of accounting for it in total scores. While there has been a lot of debate about correction formulae for yes/no tests, it seems that tests such as the VLT or the VST would also have to be systematically adjusted in light of the present findings.

In terms of form-meaning link knowledge representation, the findings did not show that one format performed considerably better than the others. The error in measurement was consistent at around 25%, with only the MC format performing marginally better. This leaves the first research question unanswered as no one format emerged as best representing meaning recall knowledge. The ambiguity of the results might indeed be due to meaning recall being a distinct type or degree of strength of word knowledge, which means the constructs underlying the various formats are slightly different. Schmitt (2014) states that "size tests based on meaning recognition will likely produce higher size estimates than those based on form recall item formats. Testers thus need to consider which form-meaning level they wish to use, and explicitly state to the end user how this should guide their score interpretations" (p. 943).

However, the results shed light on the workings of individual formats and their systematicity in over- and underrepresenting this knowledge type, which means that a link between the item score and the verified knowledge could be established even though there is no one-to-one interpretation of scores. From this viewpoint, the MC format could be taken as the favourable format for vocabulary test design based on these findings as it not only outperformed the other formats slightly as regards to representing the criterion measure but also showed the highest systematicity in its mismatches. This, in turn, would allow for methodical score correction and a more precise score interpretation. Since recognition formats generally also have the advantage of being completed faster by candidates than recall formats, and thus allow for testing a greater number of targets within a certain amount of time, a case for the MC format could be made both from an empirical as well as a practical standpoint.

The hypothesized advantage of recall formats of reducing guessing probabilities could not be fully confirmed in the results of the present study. Rather, these formats (DEF and CON) were found to underrepresent word meaning knowledge, or rather to both under- and overestimate word knowledge to almost equal amounts, which renders them problematic options for test construction.

The second research question this study attempted to answer was whether the item formats in question yielded any additional valuable information about the word knowledge of participants, that is, could they be interpreted as showing word knowledge beyond just the form-meaning link knowledge. A comparison of test item scores with the respective items testing the derivative and collocational knowledge of candidates pertaining to these targets showed that test items could not indicate derivative or collocational knowledge to any great degree. This could underline the need for additional derivation or collocation measures to be included in a test battery that aims at making claims about candidates' knowledge in these word knowledge areas.

The recall formats outperformed the recognition formats in representing collocational knowledge. This could be explained by current theories of lexical development, which maintain that collocational knowledge of a word is acquired at a later stage (Schmitt, 2010), in this case corresponding more with the higher degree of strength of word meaning knowledge elicited by the more challenging recall formats.

The results of a comparison between derivation test scores and meaning recall knowledge seem to suggest that it cannot be assumed that the knowledge of a word family member's meaning does imply receptive knowledge of other word family members. Candidates in this study could not consistently make the connection between derivational forms of a word and its base, even though they demonstrated knowledge of the meaning of that base word. This echoes findings by Schmitt and Zimmerman (2002) as well as those by Ward and Chuenjundaeng (2009), further rendering word family lists as sampling basis for vocabulary tests questionable. The evidence presented here appears to support the case for the use of lemma lists as the basis for vocabulary tests over word family lists. It certainly raises further doubts about test score interpretations of established tests such as the VLT or the VST, in that their knowledge estimates are overly optimistic, even leaving aside the abovementioned lack of correspondence between test item score and verified meaning recall score in all cases.

In further analyses, the three types of knowledge tested in this study were related to each other. Hypothesising from the research literature on vocabulary acquisition and development, one might postulate that different aspects of word knowledge develop at different paces. For instance, Schmitt

(2010) suggests that collocational knowledge might be later acquired than derivational and form-meaning link knowledge. Given that form-meaning link knowledge is often seen as one of the most basic forms of word knowledge, it could be hypothesised to be learned earlier than both derivational and collocational knowledge. It was therefore deemed interesting to probe whether any of these "higher" or "later acquired" knowledge types might imply knowledge of a more basic or "earlier acquired" knowledge type. Assuming the written meaning recall test was the most comprehensive measure of form-meaning link knowledge, the scores of the meaning recall test were therefore compared to the scores in the derivation test and the collocation test to explore whether there is some kind of implicational hierarchy between the word knowledge aspects. It was found that the collocation measure could indeed be regarded as the "most superior" or strongest/latest form of word knowledge of the three as knowing the collocations implied a 75% probability of knowing the word's meaning and a 69% probability of knowing its derivative forms. People demonstrating derivational knowledge of a word, were able to answer the meaning recall items in 67% of the cases. However, only 44% of the people scoring on the derivation items, also demonstrated knowledge of the collocations of the respective items. This could mean that collocation knowledge is a higher or more difficult type of word knowledge and therefore a better indicator of successful mastery of derivative knowledge. Knowing the meaning of a word implied knowledge of derivational forms in 73% of the cases, while the proportion of people who answered the meaning recall item correctly and answered the respective collocation items correctly was relatively low at 53%. This may suggest that meaning knowledge sits at the middle between the other two types of knowledge in terms of difficulty, although it seems fairly similar to derivative knowledge. This could be understood to mean that form-meaning and derivative knowledge are probably more basic aspects of word knowledge that are acquired relatively early, while knowing the collocation of a word indicates a fairly solid mastery of other word knowledge aspects.

These results are interesting both in light of vocabulary development theories and vocabulary test development. The findings give some support to the hypothesis that vocabulary learning is incremental and different knowledge aspects might develop at different rates, although the generalizability of this claim is obviously limited by the one-off nature of the study design, the sample size and the sample population. However, the results could have implications for the design of diagnostic lexical knowledge measurement tools, particularly in terms of their score interpretation and their test design in computer-adaptive test batteries.

The results of the analyses of the test item formats could be taken to mean that the agreement between the different knowledge aspects and therefore the representation of one knowledge aspect test through the score on another one is unsatisfactorily low and that therefore a test, which wants to make valid score interpretations for a number of knowledge aspects needs to be testing these aspects in question separately. However, when accepting the systematic misrepresentation of verified word knowledge by different test item formats and accounting for that, the findings of the comparisons between the three knowledge types could imply that a computer-adaptive test battery could be devised that presents candidates with collocation items first, as they predict a certain level of mastery in the other word knowledge types, and only presents them with form-meaning link items if this first threshold has not been mastered successfully. This hypothesis, of course, will need to be probed further as such a procedure might result in an underestimation of a person's form-meaning link knowledge because of their success in answering collocation items.

Also, this claim needs to be substantiated further due to the limitations of this study. For reasons of practicality the number of target items (k=36) had to be kept relatively small. Given the comprehensive nature of the investigation into different word knowledge aspects, however, 36 items seemed a relatively solid sample size compared to other vocabulary test studies (e.g. Paul et al., 1990). Further research would also need to extend the study to populations of different L1 backgrounds and more heterogeneous groups of age and language proficiency. Moreover, there may be other aspects of word knowledge that provide a clearer hierarchical relationship between each other, which were not investigated in this study. However, even with only the three aspects included and despite careful consideration in the study design, any cross-contamination or influence of the different tests on each other cannot be ruled out with absolute certainty.

3.5.Summary

The study presented in this chapter has explored the usefulness of different item formats for vocabulary tests. Starting from the assumption that meaning recall formats are too impractical for large scale use, an alternative item format was searched for that represented this type of form-meaning link knowledge authentically required by readers. A comparison of one form recognition, one meaning recognition and two form recall item types with a criterion meaning recall measure thereby found that all formats represented meaning recall knowledge similarly well, but all with an unsatisfactory error in measurement of roughly 25% and behaving very differently individually. The MC format, though not free of flaws, was suggested as the most promising of these for its systematicity in overestimating scores, which could be methodically adjusted on the basis of the findings. Also, the study found that other aspects of word knowledge, such as collocational and derivational knowledge, are only partially represented by these form-meaning link items. Collocation knowledge, however, was found to imply or predict a certain level of mastery of form-meaning link knowledge and derivational knowledge, which could be exploited in computer-adaptive test batteries of lexical knowledge tests. Lastly, the results were taken to make a case for the lemma as a counting and sampling unit for vocabulary tests as the assumed relationship between meaning knowledge of several members of a word family have to be doubted. The implications of using lemma lists for item sampling and the resulting issues of sampling rate and target population size are therefore explored in the next chapter.

4. Item sampling

This chapter discusses issues related to item sampling in vocabulary tests. It will do this by exploring two major concerns in this area: (1) the counting unit, and (2) the sampling principle of frequency and issues of sampling rate. The chapter will present two studies that probe the notion of frequency as a clustering factor and attempts to find an improved sampling rate through corpus analyses.

4.1.Counting unit

Before sampling rates can be discussed, the counting unit of a vocabulary test needs to be problematized and defined. Most vocabulary tests to date have been sampling based on frequency with the word family as the counting unit. The VLT (Schmitt et al., 2001), the VST (Nation & Beglar, 2007), the Eurocentres Yes/No test (Meara, 1992) as well as the CATSS (Laufer & Goldstein, 2004) are all examples of word family-based vocabulary tests. Even very recently developed tests, such as the Lexical Test for Advanced Learners of English (Lemhoefer & Broersma, 2012), the New Vocabulary Levels Test (Kramer & McLean, 2015), the Listening Vocabulary Levels Test (McLean et al., 2015) or the Picture Vocabulary Size Test (Nation & Anthony, 2016), work with this counting unit. The assumption behind using this counting unit is that the test score on one representative of a particular word family can be inferred to represent knowledge of not only that particular word item, but also all members of its respective word family. If a candidate knows one word family member, it is taken for granted that they also know the other word family members, at least to the extent that they can connect the word family members in their lexicon, which supposedly aids understanding, particularly in language reception. Using word families as counting unit therefore theoretically holds great potential for practicality and generalizing test scores: Given that each word family has between 4 and 6 members on average (Nation, 2006), few individual words need to be tested to infer knowledge of a relatively large number of individual lexical items.

However, this notion has been contested by several research studies, not least the one presented in Chapter 3 of this thesis. The research presented in Chapter 3 found that EFL learners who knew the meaning of a base word managed to connect its derivative forms (i.e., other word family members) to that base word in only about 73% of the cases. Schmitt and Zimmerman (2002) showed that EFL learners were able to produce the four classes of word family members only for about 19% of the words they were tested on. Ward and Chuenjundaeng (2009) concluded from their suffix knowledge study with Thai EFL learners that their findings "contradict the assumption that knowledge of headwords implies knowledge of word families, at least with lower-level students from non-Latinate L1 [first language] backgrounds" (p. 465). There is also psycholinguistic evidence that indicates that second language (L2) processing relies less on morphological decomposition than L1 processing and that links between word family members might thus not be very strong in L2 learners' mental lexicons (Silva & Clahsen, 2008). While L2 learners clearly have some knowledge of the relationships between word family members, this level of knowledge appears to be much less robust than a word family–based vocabulary test development and score interpretation would acknowledge. Neither productively, nor receptively have learners been shown to live up to the theoretical expectations. The fact that learners can demonstrate knowledge of the meaning of one word family member does not imply they also know its derivative forms. As outlined in Chapter 2 of this thesis, the word family is also misrepresenting the nature of language and lexis as it falls short of accounting for multi-word expressions and other formulaic sequences which are lexical in nature and ubiquitous in language use (Schmitt, 2010).

The concept of the word family as a counting unit for sampling in vocabulary tests can therefore not be maintained. Looking for alternatives, taking each word family member, including inflectional forms, as an individual item to sample from seems also rather unhelpful as that severely restricts the generalizability of results. A very large number of words would need to be tested to arrive at meaningful estimates. Also, it appears unlikely that one would want to test several inflectional forms of a word in one particular test, even though, in a Sinclairian fashion, this might be desirable as each form is indeed characterized by different properties in terms of its usage and collocations (Sinclair, 2004).

The lemma, defined as the base word and its inflections (Nation & Waring, 1997), might therefore offer a reasonable balance between clustering words together to some degree while at the same time maintaining interpretability of scores. Since the representatives of a lemma differ only in grammatical form rather than lexicosemantic properties, at least in most cases, knowledge of one lemma representative would most likely imply knowledge of the other lemma members. Using lemmata as counting units would therefore enhance interpretability of scores so that we would have a clearer idea of what a correct answer on an item does and does not mean. Additionally, multi-word expressions could be integrated into lemmatized lists, as has been demonstrated by Martinez and Schmitt (Schmitt, 2012). More recently, this debate about the counting unit has attracted some attention in the vocabulary research community. Pinchbeck (2016) argued convincingly that the optimal counting unit (or definition of word) may differ for different test taker groups. However, his findings seem to suggest that the lemma might be the most workable unit for most general test purposes, particularly for beginner to intermediate English language learners. McLean (2017) also puts forward evidence for adopting the lemma (or flemma, as he refers to the unit) for vocabulary testing and instruction. Also, lemmas have already been shown to hold advantages over word families in lexical diversity measurement (Treffers-Daller, Parslow, & Williams, 2016).

4.2.Frequency-based item sampling

Closely related to word family-based item sampling in vocabulary tests, is the notion of word frequency. Starting with the first publication of the Vocabulary Levels Test (Nation, 1983), sampling based on word family frequency lists has become the norm in vocabulary tests, particularly those of vocabulary size and those designed for international usage. In general, a frequency-based approach appears to make sense. Vocabulary (size) test scores need to be interpreted meaningfully in terms of what a particular level of word knowledge would allow a learner to do, and frequency levels have been established to relate to particular language tasks through coverage research. Although such coverage research is again heavily reliant on the limited notion of word families, it is now generally accepted that it does have added value in identifying the lexical demands put on learners in different language-related activities. There is no reason why such coverage research could not be updated using lemmatized word frequency lists, as has been hinted at by Brezina and Gablasova (2013). As of yet, however, the replication of seminal coverage research studies using lemmata instead of word families, has only been suggested but not carried out (Schmitt, Cobb, Horst, & Schmitt, 2016). In addition, little research has looked into the usefulness of frequency as a sampling criterion across different frequency levels, which might be variable. Also, no research to date has taken an empirically-based approach to clustering, but has instead mostly relied on the pragmatic decision to group items together into bands of 1,000 word families.

The design rationale of the VLT, which was strongly guided by the coverage research available at the time, was critical in the adoption of this approach. At the time of its initial publication, about 2,000 word families were estimated to be enough to engage in daily conversation, 3,000 word families were deemed sufficient to access authentic reading, while 5,000 word families were thought to enable independent reading and 10,000 word families advanced usage in several skills and domains (Schonell, Meddleton, & Shaw, 1956).

Although there is still a dearth of research on lexical requirements for language production, latest research has corroborated some of the figures for reception. Van Zeeland and Schmitt (2013) found in their study that around 2,000-3,000 word families are needed for conversational listening, adopting a 95% coverage threshold for comprehension. Adolphs and Schmitt (2003) claimed that about the same amount of word families enables to engage in basic daily conversation. Webb and Rodgers (2009) demonstrated that learners require about 3,000 word families to watch and largely understand movies and television programs, thereby confirming the importance of high frequency vocabulary. The figures for written reception, i.e. reading, however, had to be revised in light of recent findings. In terms of lexical demands for reading, Nation (2006) claimed that 8,000-9,000 word families were needed for fluent reading. Schmitt and Schmitt (2014), for this reason, also argue for the teaching, and therefore testing, of this mid-frequency vocabulary of between 3,000 and 9,000 word families. While 3,000 words might be enough to arrive at reasonable comprehension of and initial access to authentic listening, viewing and reading texts, knowledge of these additional word families would certainly make any of these experiences less strenuous and thus more enjoyable (Schmitt & Schmitt, 2014). Updated measurement instruments that assess lexical knowledge at these newly identified crucial frequency levels would therefore be desirable. However, these tools have yet to be designed and demonstrated to yield similarly valid and reliable results as established vocabulary tests.

Harking back to the above discussion, the question remains, however, whether lemmatized coverage research would corroborate the findings of these frequency levels being linked to successful language use. Nonetheless, coverage research does provide one promising way to identify a reasonable and empirically grounded population size from which vocabulary items should be sampled, while at the same time allowing for meaningful score interpretation by linking results to employability in language skills. This is, however, contingent on frequency being maintained as a useful ranking and clustering factor of vocabulary items.

4.3.Frequency as clustering factor

Frequency is generally assumed to be a key factor in language learning (Ellis, 2002). As such, it is also taken to be a relatively strong indicator of word difficulty and therefore a useful clustering factor in item sampling. The reasoning behind this is that vocabulary learning broadly follows a frequency order: the more frequent a word occurs in discourse, the more important it is for language use, the earlier it is learnt. This rationale is so influential that it is often employed in vocabulary test validation in that a frequency-based test is expected to show decreasing average facility values across frequency levels.

Milton (2009) thus states that "the importance of frequency in vocabulary learning is as near to a fact as it is possible to get in L2 acquisition" (p. 242).

While this may hold true to a large degree, frequency models have themselves never been fully validated (Brown, 2012). Schmitt and Schmitt (2014) argue that we need to reassess the notion of word frequency in relation to teaching and testing value. While they argue in their paper for a revaluation of the so-called mid-frequency levels (3,000-9,000 word families), it also clearly emerges that they see 9,000 word families as the cut-off to low frequency, a point beyond which it seems frequency becomes a rather arbitrary concept that is very much domain- and corpus-dependent.

Frequency might therefore be a good clustering factor and sampling criterion at the higher end of the spectrum, with the most frequent 2,000 words perhaps being almost identical across word frequency lists extracted from various corpora, but might be less useful and less powerful towards the lower end of the frequency continuum. While the most frequent words in any corpus might be almost identical or at least have considerable overlap with the most frequent words in any other corpus, there might be a particular point along the frequency continuum where the frequency level of a particular word becomes a mere artefact of the employed corpus. Sorell (2013), for instance, found that there is considerable overlap between word lists created from different 20 million word corpora up to the mid-frequency bands. In other terms, a word that is among the most frequent 1,000 word families in the COCA is very likely to be among the most frequent 1,000 word families in the BNC. However, a word from the 15,000 word frequency level in the BNC might be at 7,000 or at 20,000 in a COCA-based word frequency list. The aim of the study presented in this chapter is to determine whether there is such a point or band on the frequency continuum at which the frequency level of a word becomes a function of a corpus and therefore relatively arbitrary, and where this point or band might be. If the hypothesis of such a posited threshold were to be confirmed, this could potentially inform and guide item sampling for vocabulary tests.

Several studies, mostly based on vocabulary size test scores, have already hinted at such a threshold. Aizawa (2006) tested 350 Japanese EFL learners on a Yes/No test on items from the JACET8000 list (JACET, 2003) and examined their knowledge profiles in terms of frequency bands. The frequency model functioned well, showing a stairstep decline in facility values as the frequency bands got lower, but only for the four most frequent bands. Aizawa thus claimed that frequency band distinctions beyond 4,000 words are relatively uninformative.

Similarly, Milton (2007) found in his study of 227 Greek EFL learners' performances on the X_Lex test that the frequency model worked well at an overall group level in distinguishing the first four of five frequency levels. After this threshold, however, the differences in facility values seem minimal. In addition, Milton analysed the individual profiles of learners and found that around 40% of learners' scores did not follow the predicted frequency model, indicating that more complex factors are at play, particularly at high frequency levels. Brown (2012), in his replication of Milton (2007) in a small-scale study in Japan using a 120 item Yes/No test with words from the JACET8000 list, found that the frequency model worked better for this group of learners than claimed by Milton.

In a different vein, Beglar's (2010) validation study of the VST could also be taken as an indicator of frequency's diminishing power as a clustering factor the further down one goes on the frequency spectrum. The uneven profiles he identified on a group level could suggest either flawed test items (which is how he explains the unexpectedly high facility value in the 8K band) or the idea that frequency is indeed less powerful as a predictor of difficulty after a particular threshold, for instance the cut-off between high- and mid- or low-frequency.

The findings of the frequency effect attenuating beyond the most frequent levels are also in line with the relative importance of these levels in terms of discourse coverage. Davies (cited in Schmitt & Schmitt, 2014) showed that beyond the first five most frequent 1,000 levels, each further levels only adds minimally, i.e. less than 1% to the coverage of the texts in the COCA. Although this result echoes earlier findings by Nation (2006), neither has been taken to question the frequency based sampling and the sampling rates of vocabulary tests to date. If it is the high frequency vocabulary that does the most work and is therefore potentially the most important for learners to master, there might be an argument for homing in on these levels in vocabulary tests instead of treating all frequency levels equally in terms of sampling.

If a decision regarding a suitable counting unit (e.g. lemma) and sampling criterion (e.g. frequency) has been made, there still remains a question about a feasible sampling rate. Practicality concerns need to be balanced with concerns for content validity in terms of adequate and sufficient sampling from a test construct in order to enable meaningful inferences from test scores. In vocabulary tests, even if they all operate with the same counting unit (mostly word families) and the same sampling criterion (mostly frequency), sampling rates differ considerably.

In different variations of Yes/No Checklist tests, up to 10 items are sampled per frequency level, although this is difficult to ascertain with so many different versions and different sampling rationales available (Beeckmans et al., 2001). For a test that is as quick and easy to administer, this rate is surprisingly low. In the VLT, Schmitt et al. (2001) have shown to improve the robustness of initial VLT versions when increasing the number of items per frequency band from 18 to 30. The VST, however, has taken the sampling rate down to 10 items per frequency level, with any one word or item representing 100 other items through the score multiplication suggested by its authors. Beglar (2010) suggests that this rate may be enough, but Gyllstad, Vilkaité, and Schmitt (2015) convincingly argue against this based on their findings. More recently, another version of the VST has been made available on Nation's website featuring 20K frequency levels but only 5 items per frequency bands. Any one item representing 200 other items, however, can hardly be justified in terms of content validity and meaningful score interpretation. In any case, it emerges clearly that there is no consensus as to how many items should be sampled from a given frequency band to make for a valid instrument.

With these crucial yet unresolved issues identified, this chapter presents two approaches to attempts to address the following research questions:

1. What is a feasible and empirically principled sample population of vocabulary items to be tested in a diagnostic vocabulary test?
2. What is the best way to group these items together in order to sample from them for a vocabulary test to allow both feasible and meaningful score interpretation?

The first approach to inform decisions on these issues will be coverage-based, exploring the coverage level of different frequency-based word lists in different corpora. The second approach will be using test scores, comparing scores on a reading comprehension test and scores on a lemmatized, frequency-based vocabulary test.

4.4. Informing item sampling through coverage figures

To answer the two proposed research questions, the coverage provided by lemmatised frequency lists of different corpora was compared. For this, frequency lists of the following four corpora of English were extracted. The Corpus of Contemporary American English (COCA) (Davies, 2008-) was chosen as the reference corpus as it provides the largest, most up-to-date, systematic collection of texts in English from a variety of genres, including spoken language in the globally most prominent variation of English. This purchasable frequency list was compared to three frequency lists extracted from the respective corpora via the platform Sketch Engine (Kilgarriff et al., 2014): one for the British National Corpus (BNC), as it is one of the most researched and largest corpora of English and in many senses the British counterpart to the COCA; one for the enTenTen Corpus as it is one of the most up-to-date British English corpora and therefore a more recent linguistic reference point than the BNC; and one for the BROWN corpus, an influential corpus still used in current corpus research for comparative purposes (Brezina & Gablasova, 2015).

The first comparison is cumulative in nature, i.e. it describes how many lemmas are shared in both lists up to a respective frequency level, taking all higher-frequency levels into consideration, rather than just comparing the sections of the respective individual 1K frequency bands. For instance, the percentage displayed for the 5K level denotes the amount of overlap between two lists from 0-5,000 lemmas rather than from 4,001-5,000 lemmas. Comparing the COCA reference list to the various lists reveals that there is most overlap with the enTenTen list. This is hardly a surprise as this is the most recent of the lists and therefore likely to resemble another list of contemporary English. Unexpectedly, the oldest of the lists (BROWN) shows the least overlap of lemmas in the lists across the frequency rankings. The BNC list most closely resembles the average overlap between the different lists with the COCA reference list. However, even in the rather atypical curve of the comparison with the enTenTen list, there is a steady decline in shared lemmas across all lists after the first 5K, indicating that indeed the inclusion of particular lemmas becomes more and more corpus-dependent the lower the frequency level we look at on the frequency cline. This is even more salient given that most of the overlap will be provided by the function words, which are generally not part of the item sampling pool of vocabulary tests anyway. Nevertheless, the comparison of lists also revealed that there is no one particular cut-off point at which the overlap between lists drops suddenly, which makes a definite decision about the remit of the sampling population that is purely empirically-motivated difficult. However, after the 8,000 band, the average overlap falls below 75%, as can be seen in the figure below.

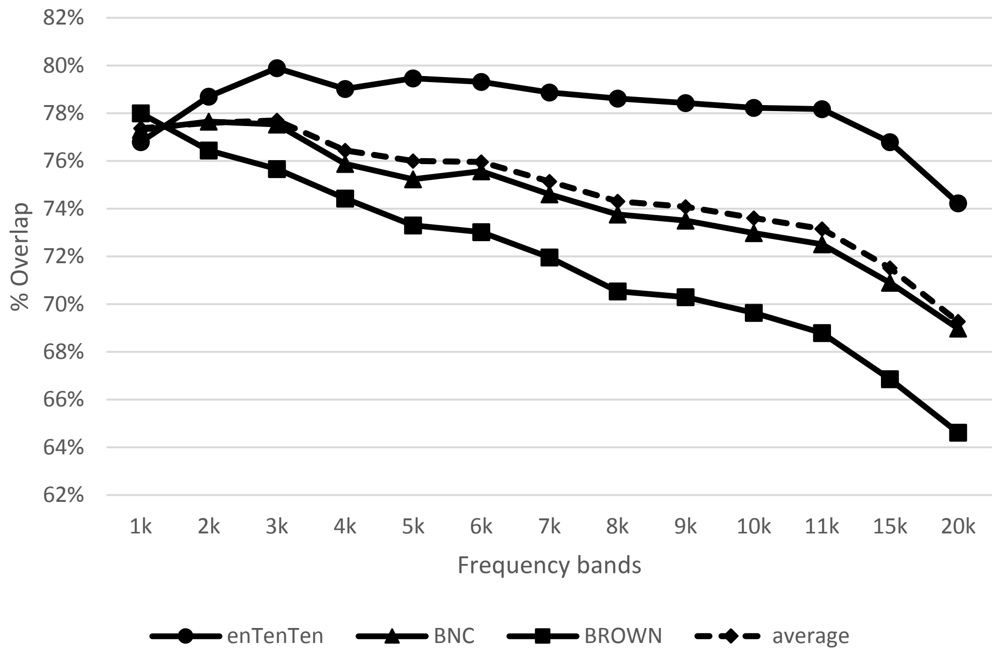


Figure 17: Overlap of lemmas in different frequency lists vs. the COCA reference list

In trying to identify the total population of items from which to sample for a vocabulary test, it would be ideal to take the suggested coverage thresholds of 95% or 98% (Nation, 2006) to decide on the cut-off along the frequency continuum. Since it has been established that the overlap between lists may not be ideal as sole criterion, this could potentially guide the determination of the size of the target population. In the reference corpus COCA, it appears that this threshold of 95% is not realistically attainable. Even the most frequent 60K lemmas only provide about 92% coverage in total, which is somewhat at odds with previous estimates that 9,000 word families with about 5 family members on average (Nation, 2006) provide about 98% coverage (Nation, 2006) (9,000*5=45,000). The addition of 3-5% proper nouns, as estimated by Davies (cited in Schmitt & Schmitt, 2014), however, means that this critical value could almost be reached, even when applied to a much larger and more diverse corpus than the one used for the establishment of the critical value. On average, however, 10,000 lemmas provide about 93% coverage of a respective corpus as Table 27 displays.

Table 27: Coverage of lemmatized frequency lists in different corpora

	COCA	BNC	Brown	enTenTen	Average	Average (-COCA)
10,000	88.8	94.6	95.1	92.3	92.70	94.00
15,000	90.2	96.3	97.1	94.2	94.45	95.87
20,000	90.9	97.2	98.2	95.2	95.38	96.87

Nonetheless, this finding either calls into question the posited 98% threshold value or the word-family based vocabulary size estimates put forward based on this figure.

Table 28 below illustrates, however, that 10K of lemmas already approach 90% coverage and that the subsequent frequency bands add only minimally to the total coverage. 10,000 further lemmas only add 2.15% of coverage. This questions whether the inclusion of such a large additional sample into the total population can actually be warranted from a practicality perspective. For purpose of the present vocabulary test, it seems therefore that the data suggests to limit the item sampling to the first most frequent 10,000 lemmas of English.

Table 28: Coverage of lemmas in COCA by frequency level 10K-20K

Frequency level (lemmas)	Coverage
10,000	88.77%
11,000	89.15%
12,000	89.47%
13,000	89.75%
14,000	89.99%
15,000	90.19%
16,000	90.38%
17,000	90.54%
18,000	90.68%
19,000	90.81%
20,000	90.92%

In a second step, all function words were removed from the reference COCA list, based on the categorization by the frequency list designers and the definition of function words proposed by Leech, Deuchar, and Hoogenraad (1982). Following this procedure, only nouns, verbs, adjectives and adverbs remained in the list as content words. The other lemmatized lists did not contain this word class information, which unfortunately made it impossible to perform the same reduction procedure on them. Further analyses were therefore only performed on the reference COCA frequency list.

Following this, the reference list itself was examined for the coverage their frequency-ranked lemmas provided. In line with Davies' estimate (cited in Schmitt & Schmitt, 2014), function words provided about 40% coverage of the corpus. Most of these 40% can be accounted for by the 127 function words found among the first 500 lemmas in the COCA frequency list.

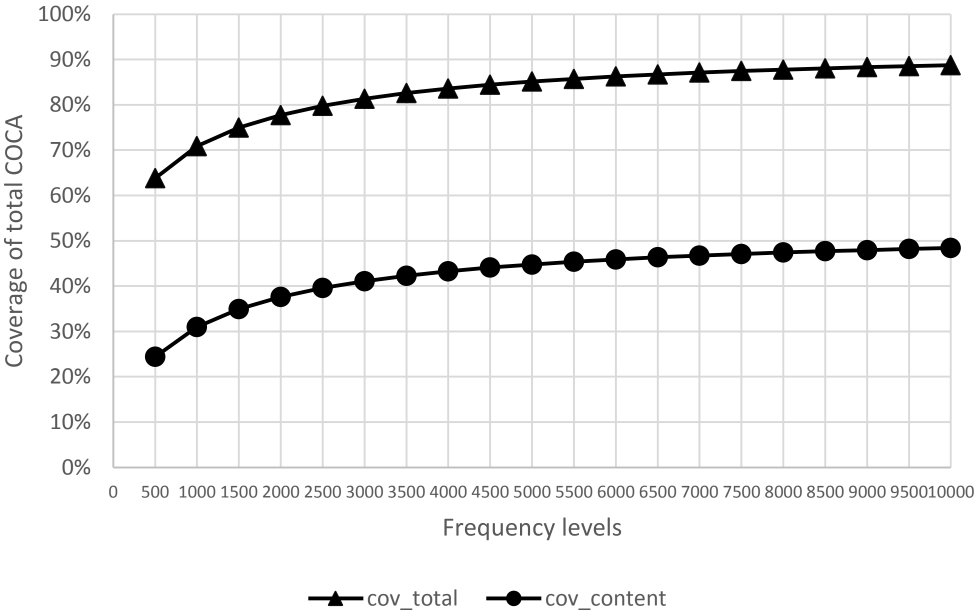


Figure 18: Coverage provided by all lemmas vs. coverage provided by content lemmas only, ranked by frequency levels

Figure 18 illustrates that vocabulary tests, in reality, sample from a pool of items that provides much less coverage than score users are led to believe. Although the first 500 lemmas in the COCA list provide about 65% coverage, only 26% of that comes from content lemmas. About 40% coverage is provided by function words and although the subsequent frequency bands provide additional coverage, the added value is relatively limited. The figure also highlights the need to break up the convenient high-frequency 1K levels into finer-grained bands as these lemmas are, based on the coverage they provide, simply more useful and important for language learners. It would thus make sense to sample more and in more detail at this end of the frequency continuum and cluster lemmas together in bigger bands towards the lower-frequency end as they are of limited use in the additional coverage they provide. Table 29 illustrates this. A reordered list of content lemmas shows that the 500 most frequent content lemmas provide 26.73% coverage in the COCA corpus. While the next three bands of 500 add a further 5.82%, 3.51%, 2.43%, 1.80%, and 1.40% coverage respectively. It also emerges that a bigger cluster of lemmas might be useful at this point from a coverage perspective, suggesting that mid-frequency vocabulary between 3K and 6K, could be split into three 1K bands. This would also be mostly in line with Schmitt and Schmitt's (2014) suggestion of a tripartite notion of high-, mid- and low-frequency vocabulary. The last group of lemmas provides very little additional coverage, which is why it could be argued that two frequency clusters, i.e. 6-8K and 8-10K, could be fine-grained enough to sample items from.

Table 29: Coverage provided by content lemmas split into frequency bands

Frequency level (of content lemmas)	Coverage provided	Coverage gain per band
500	26.73%	26.73%
1,000	32.55%	5.82%
1,500	36.07%	3.51%
2,000	38.50%	2.43%
2,500	40.30%	1.80%
3,000	41.70%	1.40%
3,500	42.81%	1.11%
4,000	43.72%	0.91%
4,500	44.48%	0.76%
5,000	45.12%	0.65%
5,500	45.69%	0.57%
6,000	46.17%	0.48%
6,500	46.59%	0.42%
7,000	46.96%	0.37%
7,500	47.29%	0.33%
8,000	47.59%	0.30%
8,500	47.86%	0.27%
9,000	48.10%	0.24%
9,500	48.32%	0.22%
10,000	48.52%	0.20%

It needs to be acknowledged at this point that any division of frequency bands or clusters will probably be arbitrary, even if it was done based on a transformation of frequencies onto a log scale. Although such a log-based Zipfian approach to assessing vocabulary size has already been employed with some success in the assessment of productive vocabulary size (e.g., Edwards & Collins, 2011) future research would have to demonstrate the usefulness of such an approach to banding in item sampling for discrete receptive vocabulary knowledge tests. Following such an approach, it would hypothetically also be possible to abandon frequency bands altogether, treat any item as a measurement point on the frequency curve and estimate a vocabulary knowledge curve from that. Despite the theoretical possibility of this, research would first have to demonstrate that this is feasible and valid, and it would further appear unlikely that practitioners would find such a heavily mathematical approach accessible and practical (Cobb, personal communication).

4.5.Informing item sampling by linking test scores to skills tests

Given that a cut-off at 10,000 lemmas emerged from the analyses of frequency lists as a reasonable and feasible population to sample from, the next step was to link vocabulary size estimates yielded by a test based on this population to scores on a reading comprehension test. If no ceiling effect in this vocabulary test was observed with learners who could demonstrate good comprehension of written texts in the reading measure, this would be further support for capping the sampling of a vocabulary test at this frequency level. In addition, it would provide further evidence to question the estimated vocabulary size requirements postulated for reading comprehension.

4.5.1. Procedure and participants

To investigate this issue, 75 intermediate EFL learners from Austria were administered a vocabulary size test based on a list of the 10,000 most frequent content lemmas and a reading measure. The participants were all students of English language and literature at an Austrian university.

4.5.2. The reading measure

For the purpose of this investigation, an Aptis reading test was selected as measure for reading ability. Aptis is a multilevel language skill test suite, professionally developed and administered by the British Council (O'Sullivan, 2015). Developed for learners aged 16+, it is designed to measure reading ability up to the C1 level on the Common European Framework of Reference (Council of Europe, 2001). Being a multilevel test, it includes a range of items from different levels and reports results both on a numerical scale (ranging from 0–50) and as a CEFR level. A sample of a reading suite, provided by the

British Council, was administered to the participants on a computer. In this 30-minute reading test, participants were asked to complete four tasks, each linked to one CEFR level, A1 to B2. In the APTIS test, the A1 task consists of five three-option MC questions that are generally aimed at sentence comprehension. Candidates are asked to read and complete free-standing sentences of a text with the appropriate grammatical form or word (British Council, 2013). Task 2 assesses a candidate's knowledge of text cohesion by asking them to reorder jumbled sentences to form a (often narrative) text of about 100 words. The third reading task is a banked gap-fill and aims at testing short-text comprehension (~150 words). Since the Candidate Guide recommends practice readers such as Penguin Readers Level 4, which claims to be aiming at CEFR B1 level, it can be assumed that this level is also targeted here. While the first three tasks appear to focus on careful reading, the fourth task assesses a mixture of expeditious and careful reading behaviour. In this arguably most challenging of the four tasks, candidates have to read a longer text (about 750 words) and match headings to the text's paragraphs. The Aptis test developers claim that the four tasks also elicit a broad range of cognitive processes according to Khalifa and Weir's (2009) model, which has been partly confirmed by Brunfaut and McCray (2015). Example items of an Aptis reading suite can be found online (British Council, 2013).

4.5.3. The vocabulary size measure

The vocabulary size measure was designed along the model of the Vocabulary Size Test, albeit with a modification in the sampling population, i.e. a different word frequency list. The list referenced above contained the 10,000 most frequent lemmas of the COCA. From this list, 10 items per 1K frequency band were selected and turned into four-option multiple-choice items with one correct answer and three incorrect definitions or synonyms. All options were informed by definitions from language learner dictionaries and wherever possible, it was ensured that the defining vocabulary was of a higher frequency than the target word. This proved extremely difficult and at times impossible at the high frequency end of 1K and 2K lemmas, but was adhered to as best as possible throughout the test. The target words were always chosen from the middle of a particular 1K frequency band so as to make sure there was a distinct frequency difference between the item clusters. Also, the target words were selected according to part-of-speech in order to represent the word class ratio of content lemmas in each particular frequency band. This test of 100 items was then administered on a computer to the participants immediately after they had taken the reading test (see Appendix E for all items). The scores of the two tests will be compared in the following.

4.5.4. Results

No candidate scored lower than 38 out of 50 (=76%) on the reading test suite. Most candidates (37) scored the maximum in the test and were labelled as CEFR C level readers with no participant showing a lower proficiency than CEFR B2 in reading. Item level data for the reading measure was not available to the researcher.

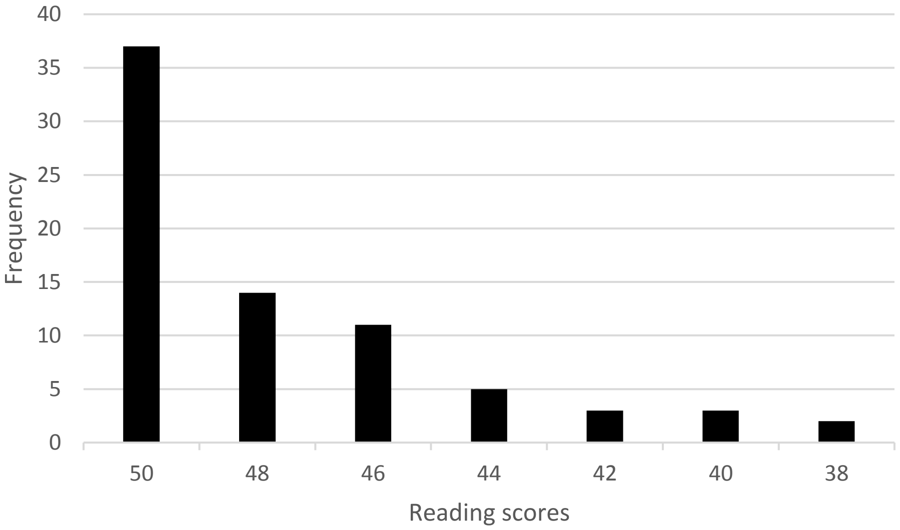


Figure 19: Frequencies of APTIS total reading scores

In terms of the vocabulary measure, Figure 20 shows that participants scored an overall mean of 81.73 (SD=9.25) of 100 items, which (according to VST reasoning) would translate into this very proficient group knowing, on average, the most frequent 8173 lemmas. This very good result and negative skew (-.3) is hardly surprising in light of the advanced language level of the group as ascertained by the Aptis reading test.

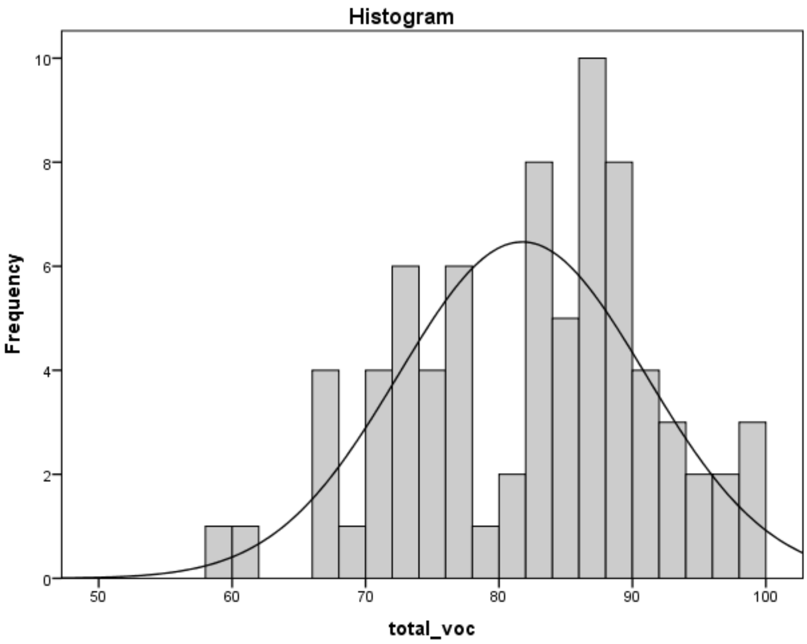


Figure 20: Histogram of total vocabulary scores

What is partly surprising, however, is that the scores of this high-level group ranged between 59 and 99 points with no participant maxing out on the vocabulary test, even though it sampled from a lemmatized list rather than a word family list. One explanation for this might be that, although the sampling criterion is different, the items themselves are not that different from a word family based test. It is at odds, however, with the notion that a word family list's sampling is so much wider, which would lead to the expectation that very proficient readers should do even better at a vocabulary test based on the 10,000 most frequent lemmas. If 8,000-9,000 word families are needed for proficient reading, as has been claimed, this would translate into a much higher figure of lemmas that need to be known than the 8173 exhibited by this proficient group of EFL readers. Given the homogenous high-proficiency nature of the participant population, the descriptive statistics in terms of the vocabulary test's reliability with a Cronbach alpha of .89 are satisfactory.

Looking at the vocabulary score profile across the lemma frequency levels, one can detect a frequency effect, albeit a very attenuated one. As can be seen in Figure 21, there is a slight drop in mean scores at 5K after the first four bands yield very similar means. The steady decline continues until 8K, at which point there is a sudden surge in mean scores, which then again decreases at 9K and slightly increases at the 10K band. It is worth noting that this profile, although based on a vocabulary test with a radically different sampling criterion (lemmas vs. word families), is very similar to what Beglar (2010) found in his validation study of the VST, particularly regarding the unexpected surge at the 8K band. The fact that the frequency profiling seems to become less predictable in the region of 8-10K could be taken as further evidence that sampling beyond this frequency band is less useful.

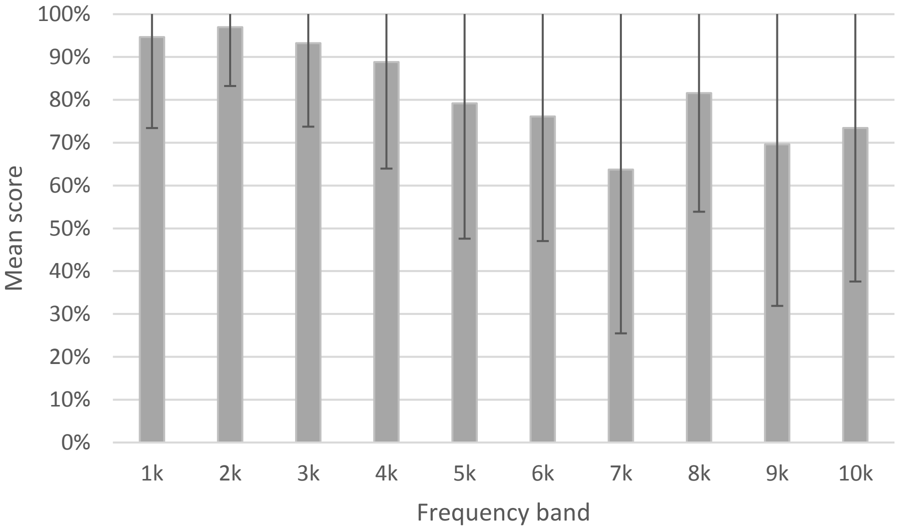


Figure 21: Vocabulary score profile across frequency bands

Plotting the reading scores against the vocabulary scores, a significant correlation between the two measures (Spearman's rho=.365, p=.001) can be seen. However, the fact that this correlation is only of medium strength is likely to be due to the ceiling effect in the reading measure and the homogenous nature of the participant sample. The scatterplot in Figure 22 nevertheless reinforces that even very advanced EFL readers show a vocabulary test score range between 69 and 98 and did therefore not manage to exhaust the lexical measure.

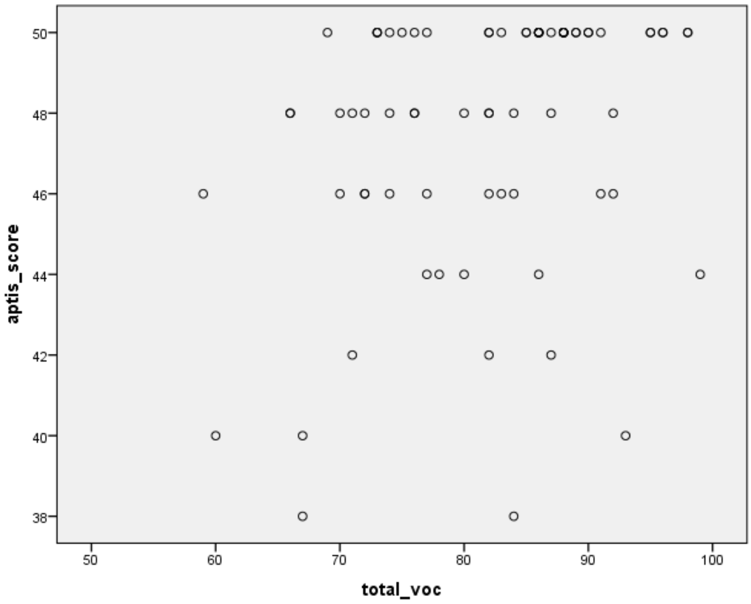


Figure 22: Scatterplot of reading scores vs vocabulary scores

4.6. Discussion

Both the results of the corpus-based coverage investigation and the analysis of proficient readers' scores on a lemma-based vocabulary test appear to point to four main outcomes. First, the lemma can be a useful counting unit for item sampling in vocabulary tests. It facilitates score interpretation and does not increase the total item sampling population considerably while at the same time retaining reasonably good interpretability in terms of coverage figures and linking test scores to the ability to perform particular tasks in the foreign language. Second, frequency is, to date, the most useful sampling criterion for vocabulary tests and although its power as a clustering factor decreases considerably along the continuum, particularly as we move into the mid-frequency bands, frequency profiles still show, even in lemma-based vocabulary test scores. Frequency-based sampling thus seems, also for lack of a better alternative, the way forward in vocabulary testing as it will allow links to lemmatized frequency-based coverage research results that appear to be in demand in the field. Third, however, as the data presented suggests that frequency's clustering power decreases, a re-evaluation and re-conceptualisation of frequency bands seems necessary. From a coverage perspective, it would appear to be diagnostically more valuable to sample in bands of 500 at the high frequency end, possibly until 3K, then move into 1K bands for the mid-frequency region, and use bigger sampling clusters at the lower-frequency end, where any further 1K only adds minimally to coverage figures. Fourth, the findings of the coverage study also suggest that 10,000 lemmas is a sufficient total population to sample from, particularly in light of concerns for practicality, as frequency levels beyond that point make up too minimal a contribution to overall coverage for their inclusion in a diagnostic vocabulary test to be warranted. This was also corroborated by the results of the study comparing reading test scores and vocabulary test scores of very proficient EFL learners as no ceiling effect in the lemmatized 10K vocabulary test could be observed. If 10,000 lemmas were enough to adequately represent the vocabulary knowledge of these advanced learners, surely this total item sampling population will also suffice to model the vocabulary knowledge of users of a diagnostic vocabulary test, who will typically be of a lower proficiency level.

For the present test design this means that items will be sampled from a lemmatized frequency list. The COCA word frequency list appears the most useful and up to date sampling basis. The diagnostic test will sample from the most frequent 10,000 content lemmas, disregarding function words. The test will attempt to operationalize a new approach to frequency band clustering and will split the first 3,000 lemmas into six bands of 500, the second 3,000 into three bands and the final 4,000 lemmas into two bands to sample from.

4.7. Summary

The chapter set out to identify a feasible and empirically principled sample population of vocabulary items to be tested in a diagnostic vocabulary test. Evidence from two studies, combining coverage and test performance perspectives, have supported a cut-off at 10,000 lemmas as a reasonable sample population for this purpose. The coverage findings when comparing word frequency lists have further indicated that a different distinction in terms of frequency bands might be useful. Based on these findings, the test presented in this thesis will operationalize a clustering of six frequency bands of 500 lemmas each for high-frequency lemmas, three 1K clusters for the mid-frequency vocabulary between the most frequent 3,000 and 6,000 lemmas and two larger clusters of 2,000 lemmas each for the two lowest frequency levels in this sample population. The sampling rate from these clusters for the computer-adaptive format will be investigated in the piloting phase of the instrument.

5. Test construction and piloting

5.1. Test specifications

Test specifications are “generative blueprints for test design” (Davidson & Lynch, 2002, p. 1) and state “what is and what is not assessed” (North, 2004, p. 78) in a test. They determine “what the test should contain” (Alderson & Cseresznyés, 2003, p. 298) and provide information on the construct, item format and, most importantly, the purpose of a test (Webb, 2006). Test specifications are generative, iterative and consensus-based tools that are required to produce one or several different forms of a test (Davidson, 2012). They are useful “to communicate to different audiences the structure and content of a test” (Webb, 2006, p. 176) and should therefore be the first step in test construction. They declare the design principles and rationales behind the test development and should guide the item writing process, the operationalization or administration and are thus also highly useful for the validation of the test.

Table 30: First draft of test specifications

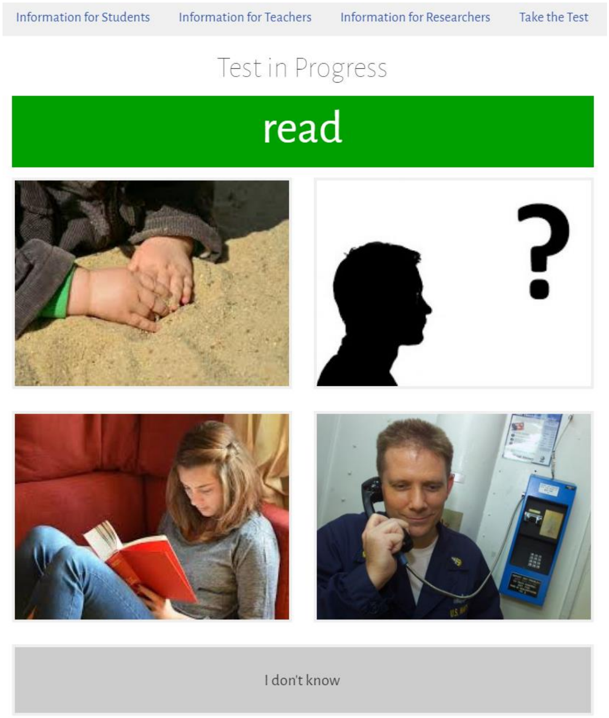
Diagnostic vocabulary test - Specifications	
General purpose	To diagnose the written receptive lexical abilities of EFL learners
Specific purpose	• To determine whether EFL learners know the written form-meaning link to the extent that it would allow employing that vocabulary knowledge for reading comprehension • To determine how well EFL learners know the form-meaning link of words from different frequency levels up to the first 10,000 content lemmas
Target language situation	International learners of EFL
Description of the test taker	All ages, but likely to be age 10 and upwards; international audience, diverse L1 backgrounds, beginner to (upper-)intermediate proficiency level
Test source	Discrete items sampled from the first 10,000 content lemmas (nouns, verbs, adjectives, adverbs) of the lemmatized COCA word frequency list (Davies, 2008-) Items will be clustered and sampled from the following frequency bands 1. 1-500 2. 501-1,000 3. 1,001-1,500 4. 1,501-2,000 5. 2,001-2,500 6. 2,501-3,000 7. 3,001-4,000 8. 4,001-5,000 9. 5,001-6,000 10. 6,001-8,000 11. 8,001-10,000
Item format	Four-option multiple choice (three distracters), target item presented in short, non-defining context, distracters either picture-based (in the first 1,500 lemmas) or text-based (synonyms, definitions) Distractors will be based on lemmas from the same frequency band that are plausible within the context of the example sentence but unrelated to the meaning of the target. Whenever an item cannot be defined in words that are of a higher frequency than the target, a picture must be used

Items per level	Approximately 10, dependent on candidate answers and the computer-adaptive algorithm
Total number of items	Computer-adaptive, dependent on candidate answers
Instructions	Target language with an example
Weighting	1 point per item
Time allowed	Untimed, but should take no longer than 30 minutes in total
Administration	Computer-delivered and scored through online website
Score reporting	In diagnostic frequency profile, split into bands and linked to information about lexical requirements of different communicative abilities (coverage research) and CEFR levels

5.2.Test construction

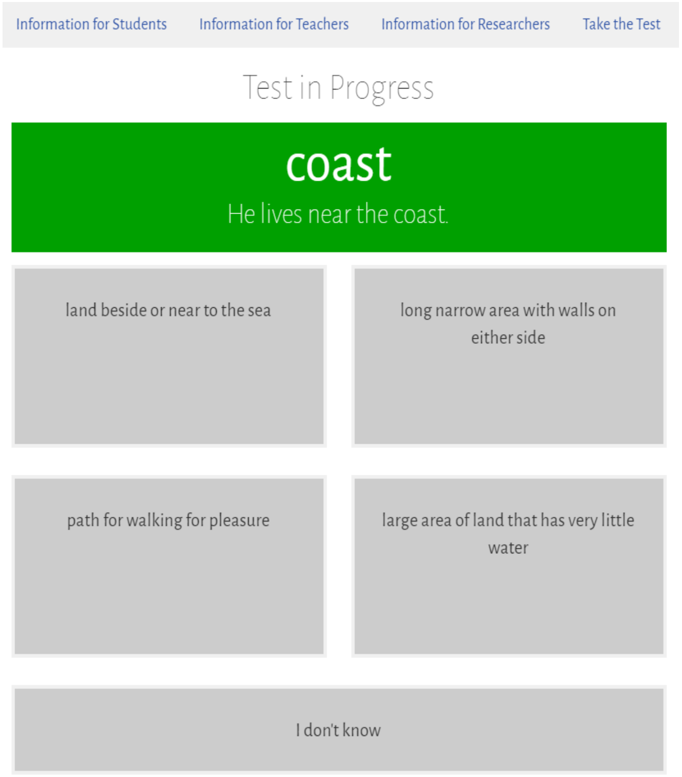
Based on these specifications, a total of 475 items were written for the 11 frequency bands. Items for the first three frequency bands were constructed using stock images under creative commons licence. For the remaining items, short definitions were used as options and distracters, which were constructed with the help of two online monolingual learner dictionaries. Examples of the items can be found in Figures 23 and 24. An "I Don't Know" option was added at the bottom of every item so that candidates could move forward in the test without forcing them to guess.

It should be pointed out that because distractors are based on lemmas from the same frequency band that are plausible within the context of the example sentence but unrelated to the meaning of the target, the current version of the test's items could be claimed to assess partial knowledge in the same way that the VST does. In other words, because the distracters are not semantically related, the items are arguably only testing a shallow depth of knowledge as they are not assessing precise knowledge. This is admittedly a weakness of the instrument in its current form. However, future iterations of the test could easily adapt to testing more precise knowledge by presenting test takers with four options that are semantically related. This would then also allow for item modification or improvement after piloting as items could be tweaked to arrive at more attractive distracters. Given the current test specifications, which follow the design principles of the VLT in this respect and guarantee a highly systematic item creation process for all items in the test, improving items is almost impossible. Any non-functioning items have to be discarded after the pilot as it would be difficult to imagine why a different non-related word would distract more than the previous one. However, a modification of this approach towards testing more precise knowledge would potentially bring to bear new issues such as subjective judgments on the relatedness of distracters and varying degrees of precision across several items. Another issue worth raising is that the test is designed for an international audience and so can currently only take L1 influences and cognates into account in a limited fashion when designing distracters. Again, though, it is an option for future test versions to have the algorithm tailor the item and/or distracter selection more to test taker characteristics. The research required to make principled decisions on these design matters, however, would be beyond the scope of this doctoral project, which is why this design approach was followed for this first version of the instrument.



read

Figure 23: Screenshot of example item for high-frequency band using pictures as options and distracters



coast

He lives near the coast.

land beside or near to the sea

long narrow area with walls on either side

path for walking for pleasure

large area of land that has very little water

Figure 24: Screenshot of example item for lower-frequency band using verbal options and distracters

After an internal review process, the items were randomly assigned to one of four static test versions for the piloting stage. First, these four test versions were piloted on 19 NS (5 took Version A, 4 Version B, 5 Version C and 5 version D), all studying for a Masters or PhD degree at a British university, to check for clarity and comprehensibility of the items. The NS were asked to complete the test in the role of a test taker and note down comments during or after the test on individual items if anything seemed unclear. Based on this first small scale pilot, 28 items were revised. 13 items were revised because of NS comments. 15 items were revised based on item analysis statistics. Items with at least one incorrect answer from a NS were inspected closely and potentially revised, particularly if they were high-frequency. Some low-frequency items, however, were not changed based on this outcome if it appeared like the item had simply not been known by the NS. Items that more than one NS answered incorrectly were revised regardless of the frequency band. After this initial mini-trial, the remaining revised items were subjected to another round of feedback and revision by an experienced vocabulary assessment specialist. This also resulted in the removal of 40 problematic items.

5.3.Trialling of item pool

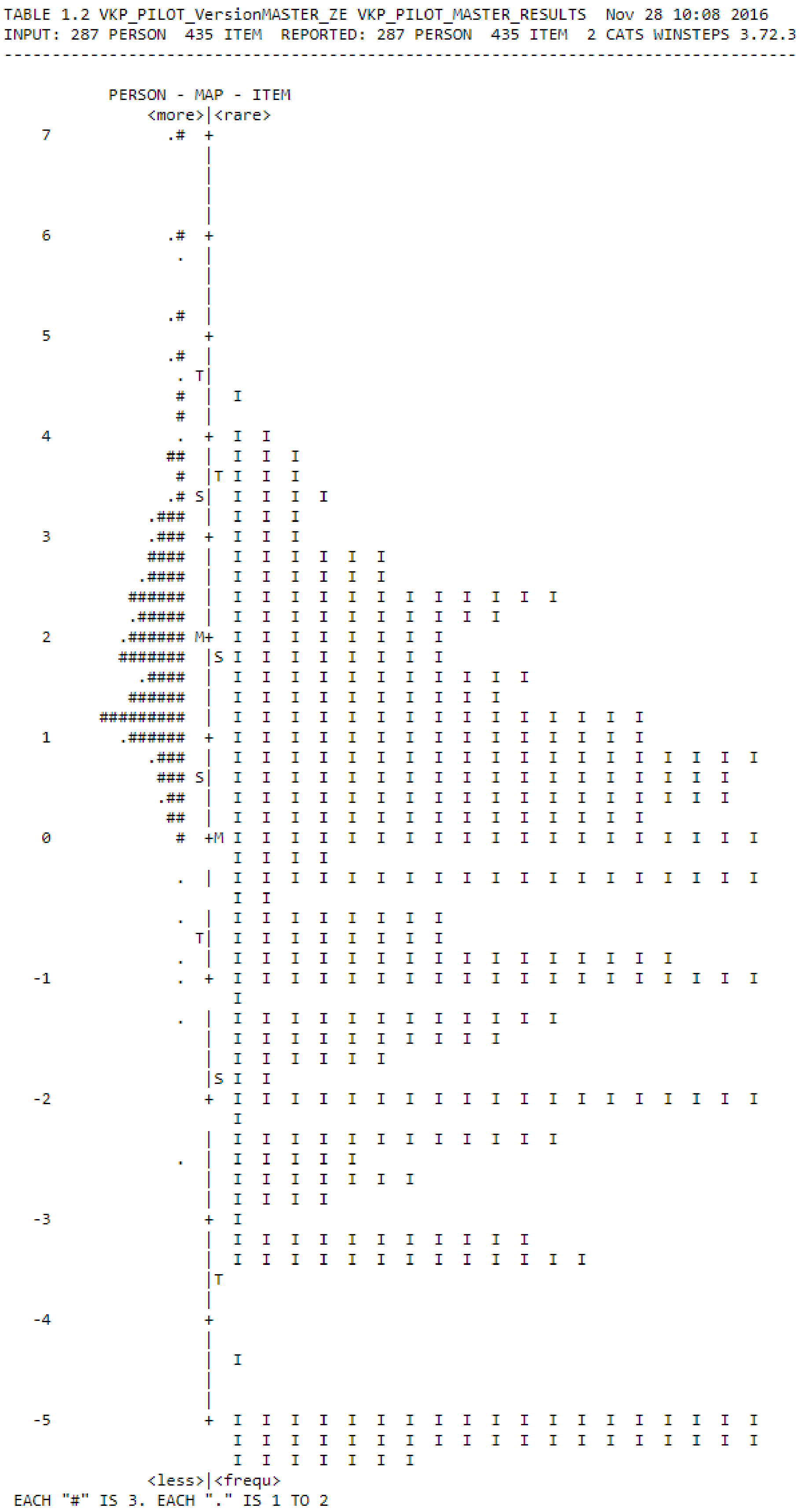
The final revised batch of items was then subjected to a large-scale international trialling with EFL learners from different L1 backgrounds. The 435 remaining items were randomly and evenly distributed across four static test versions to facilitate statistical analyses of the item functioning afterwards (see Appendix H for all target words). The four testlets were linked through 11 anchor items, one item per frequency band. Given the test specifications, there was little room for improvement of items after the trial. Distracters were unlikely to be made more distracting as the item writing guidelines specified that there were no orthographically or semantically similar options to be used as options in any one item. Hence, it was expected to reduce the item pool considerably post trial with the aim of retaining at least 25 functioning items per frequency band after the piloting.

The tests were sent out to researcher and teacher contacts all over the world. 350 participants from 20 different L1 backgrounds (Arabic, Bulgarian, Catalan, Chinese, Dutch, Fiji, Finnish, French, German, Greek, Hungarian, Indonesian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Vietnamese) participated in the item piloting. However, only 287 participants provided data useable for the item analyses. Data gathering was conducted in two waves, after the first round of data collection had not yielded satisfactory sample numbers. The participants were, on average, 22.18 years old (SD=8.56), ranging from a few 15-year-olds to one 66-year-old. Due to the nature of the international contacts and the complicated nature of obtaining parental consent from under-16-year-olds, most participants were university students. They had been learning English, on average, for 10.17 years, but the standard deviation of 5.70 years indicates a broad range of length of learning experiences. 48.8% of the sample were female, 49.8% were male.

It was left to the contacts as invigilators how they assigned the four test forms to their participant groups. At the time, this was found to be the most practical solution as the programmer was not able to implement an automated assignment of test forms in the time available. As a result, the distribution of test takers per testlet was unfortunately rather imbalanced. Testlet A was taken by 101 participants, testlet B by 81 participants, testlet C by 65 participants and testlet D only by 40 participants. This meant that there was fairly little information for about a fourth of the produced items. It therefore comes as no surprise that most of the items eliminated from the item pool based on the subsequent item analyses came from the pilot testlet D.

5.4.Trial results

The data gathered in the pilot was analysed using both classical test theory, using SPSS® 22, and item response theory, using WINSTEPS (Linacre, 2017). A first inspection of the WINSTEPS variable map showed that the population sample was relatively proficient, resulting in a slight mismatch of item difficulty and person ability.



```
TABLE 1.2 VKP_PILOT_VersionMASTER_ZE VKP_PILOT_MASTER_RESULTS  Nov 28 10:08 2016
INPUT: 287 PERSON  435 ITEM  REPORTED: 287 PERSON  435 ITEM  2 CATS WINSTEPS 3.72.3
------------------------------------------------------------------------------------

           PERSON - MAP - ITEM
               <more>|<rare>
    7            .#  +
                     |
                     |
                     |
                     |
    6            .#  +
                  .  |
                     |
                     |
                 .#  |
    5                +
                 .#  |
                  . T|
                  #  |  I
                  #  |
    4             .  +  I  I
                 ##  |  I  I  I
                  #  |T I  I  I
                 .# S|  I  I  I  I
               .###  |  I  I  I
    3          .###  +  I  I  I
               ####  |  I  I  I  I  I  I
              .####  |  I  I  I  I  I  I
             ######  |  I  I  I  I  I  I  I  I  I  I  I  I
             .#####  |  I  I  I  I  I  I  I  I  I  I
    2       .###### M+  I  I  I  I  I  I  I  I
            #######  |S I  I  I  I  I  I  I  I
              .####  |  I  I  I  I  I  I  I  I  I  I  I
             ######  |  I  I  I  I  I  I  I  I  I  I
          #########  |  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
    1       .######  +  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
               .###  |  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
                ### S|  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
                .##  |  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
                 ##  |  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
    0             #  +M I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
                        I  I  I  I
                  .  |  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
                        I  I
                  .  |  I  I  I  I  I  I  I  I
                    T|  I  I  I  I  I  I  I  I
                  .  |  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
   -1             .  +  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
                        I
                  .  |  I  I  I  I  I  I  I  I  I  I  I  I
                     |  I  I  I  I  I  I  I  I  I  I
                     |  I  I  I  I  I  I
                     |S I  I
   -2                +  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
                        I
                     |  I  I  I  I  I  I  I  I  I  I  I  I
                  .  |  I  I  I  I  I
                     |  I  I  I  I  I  I  I
                     |  I  I  I  I
   -3                +  I
                     |  I  I  I  I  I  I  I  I  I  I  I
                     |  I  I  I  I  I  I  I  I  I  I  I  I  I
                     |T
                     |
   -4                +
                     |
                     |  I
                     |
                     |
   -5                +  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
                        I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I  I
                        I  I  I  I  I  I  I
               <less>|<frequ>
 EACH "#" IS 3. EACH "." IS 1 TO 2
```

Figure 25: Item pilot variable map

Given this, item difficulty estimates, and facility values respectively, of individual items were not considered as elimination criterion. High-frequency items that had been answered correctly by more than 95% were still retained even though they would be generally removed in, for instance, achievement scenarios as carrying little useful measurement information.

Item quality was thus primarily determined by the Infit MeanSquare value in the IRT analysis, and the Corrected-Item-Total-Correlation (CITC) value in the CTT analysis. Given the parameters cited in the literature (Green, 2013), items were retained if showing Infit MnSq values between the range of .75 and 1.33. Item with values outside these parameters were considered unproductive for measurement as they were behaving too predictably or unpredictably. Green (2013) maintains that Infit MnSq values outside these critical values are a greater threat to measurement than Outfit MnSqs, which is why the former was focused on in the item selection process. The relatively strict parameters were chosen because the item pool from the pilot was large enough to apply stringent selection criteria. In the CTT item analysis, a CITC value below .25 was adopted as critical value. Items with values below this generally do not discriminate well (Green, 2013) and are thus less useful for measurement purposes. However, at the high-frequency bands, the criterion had to be softened slightly so that the minimal number of items could be retained for the pool. These items were so easy for the sample that they could not be expected to discriminate very well.

Applying the criteria, 138 items were removed from the item pool, or rather retired, available for possible re-inclusion in the future after further trialling (see above for sampling imbalance across testlets). A total of 296 items was retained, with at least 25 items per frequency band in the item pool. Table 31 illustrates how many items were removed and retained at each level (see Appendix G for full IRT results).

Table 31: Item pool after piloting

Band	removed IRT	+ removed CTT	Items remaining
1	7	7	**26**
2	4	8	**27**
3	6	6	**25**
4	6	12	**26**
5	6	8	**27**
6	3	10	**27**
7	3	6	**27**
8	4	8	**28**
9	6	7	**25**
10	2	8	**29**
11	2	9	**29**

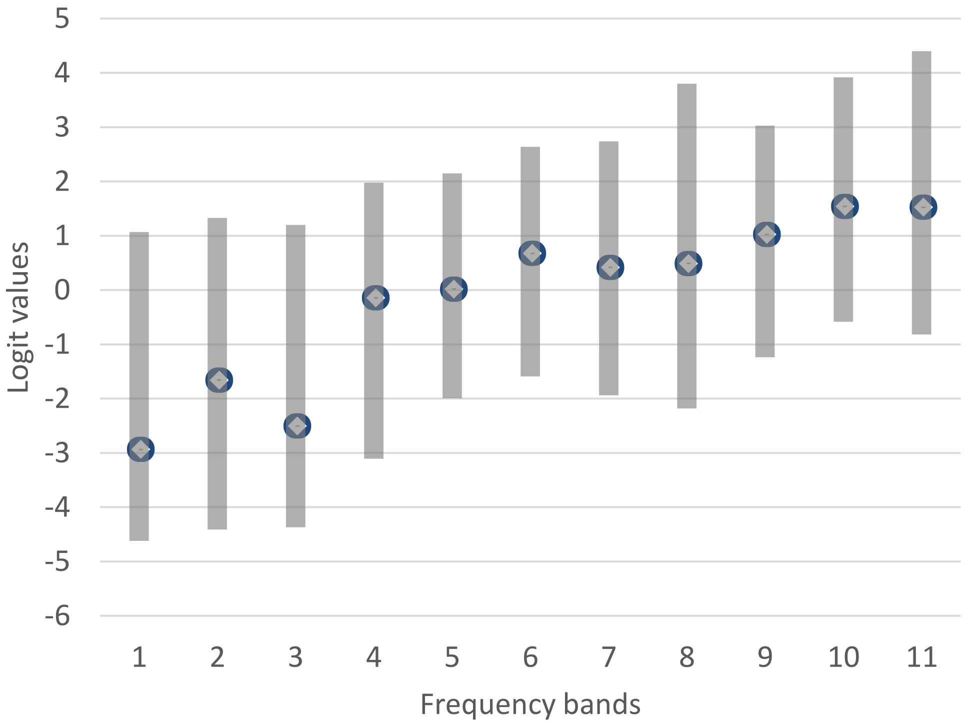


Figure 26: Item piloting logit values per frequency band

When plotting the logit values of the remaining items per frequency level (Figure 26), there is no clear stair-step profile in terms of difficulty means and ranges visible, as is usually expected in a frequency based vocabulary test. Although there seems a trend of increasing logit value ranges and means across the frequency bands, the increase is by no means linear or very distinct from the high-frequency to the low-frequency bands. There is a clearer pattern when collapsing the first six bands into three so that the bands consist of steps of 1,000 rather than 500. It seems that the fine-grained banding with this fairly proficient group of learners does not yield a distinct frequency profile in terms of item difficulty. Instead, band 3, for instance, seems to contain easier items than band 2. However, it is assumed that the frequency effect could come out more clearly, even with these fine-grained bands, if a more heterogeneous sample had been involved. This is partly because it proved challenging to find low-level EFL learners aged 16 or over. At that age, most learners in learning institutions, which was the primary way of recruiting participants, appear to have reached a proficiency level and vocabulary breadth beyond what could be measured distinctly with the first three or four frequency bands in this test.

Despite the best of the researcher's efforts to administer the test more widely and internationally, this sampling bias could not be avoided. While a limitation of the item pool in its current form, this will be addressed before the test launch by administering it to younger, lower-level learners to see if the frequency profile is more in line with previous vocabulary test research when using a sample with heterogeneous proficiency levels. Nevertheless, these 296 items were found to satisfy the psychometric standards for the purpose of the test at this stage so that they could now form the item pool from which the computer-adaptive test could sample.

6. Comparison of computer-adaptive test algorithms

This chapter discusses issues related to implementing a vocabulary test in a computer-adaptive environment. It will first review some general features of computer-adaptive tests and point to advantages of computerized testing to argue for designing such tests. The chapter will then present a study that compares two different approaches to computer-adaptive test design to evaluate which yields more robust and representative test scores.

6.1.Issues in computer-adaptive testing

The present test is conceptualised as a computer-adaptive test (CAT). This holds several advantages over designing it as a paper-pencil tool which have been outlined by the research literature. One example of such an advantage lies in improved item sampling as a computer-adaptive test can select and adjust items based on the test-takers' level of ability (Chapelle & Douglas, 2006), resulting in a more informative report of test-taker abilities. Given that the item sampling rates of existing vocabulary measures have been shown to be problematic (Gyllstad, Vilkaité, & Schmitt, 2015), a computer-adaptive test requires fewer items than a traditional paper-pencil test to determine a candidate's level of lexical knowledge, thus potentially increasing the content validity of the test. Tseng (2016) maintains that "CAT adopts a dynamic, adaptive item selection procedure to optimally target the interim ability estimate and reach the convergence, resulting in a shorter, putatively more efficient test-taking process" (p.1). Tseng, in his study, showed that the amount of items required for an accurate vocabulary size estimate could be reduced significantly through the use of computer-adaptive testing. His findings show that depending on the reduction procedure, a computer-adaptive test only required about a third of the items of the item bank to produce comparable estimates to those based on the entire item bank. Tseng's study in the Taiwanese context uses IRT-based item calibration to achieve this, employing a prescribed national curriculum-based wordlist. While such a difficulty-based approach certainly enhances reliability and measurement efficiency (Thissen, 2000) in such a relatively narrow context, it may be problematic for test that are geared towards a more general, heterogeneous and international test taker population. For instance, a vocabulary item that is Rasch-scaled at a particular difficulty level by a Taiwanese learner group, may differ considerably in its logit value if tested on a German or Swedish L1 learner group. Hence, the present project will use the logit values of the items only as secondary information in the item sampling process and primarily rely on frequency banding for item selection, at least until a large enough amount of solid data has been gathered to determine the difficulty of items for a diverse population with greater certainty than is currently possible.

Also, depending on the computer-adaptive algorithm, CAT may avoid presenting candidates with items that are too challenging and thus potentially demotivating (Tung, 1986). Since item sampling in the present vocabulary test development project will follow word frequency bands, a computer-adaptive test allows sampling in different rates from individual frequency bands to tailor exactly to the lexical needs of specific learners while at the same time providing a much more detailed inference base for any score user. Learners can therefore receive detailed feedback in the form of a graphic profile of their lexical resources rather than an overall score from a less useful one-size-fits-all paper-and-pencil test.

Another advantage of implementing computer-adaptive tests is that test instructions are presented consistently and uniformly, providing for optional but standardized help screens, to ensure fairness and comparability across all test takers (Chapelle & Douglas, 2006). Rapid automated scoring of answers through the computer also entails that feedback for users is detailed and immediate, which has been highlighted as one of the key characteristics and demands in diagnostic testing (Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja, 2015).

Chapelle and Douglas (2006) further state that computer-adaptive tests offer the option of incorporating multimodal input, which is certainly promising for the area of vocabulary assessment as it provides opportunities to replace traditional definitions with pictures, sounds, graphics interchange formats (GIFs) or even short video clips. While the use of sounds and animated pictures is beyond the remit of the present doctoral project, it offers interesting avenues to explore in further research. The present project does, however, make use of pictures as outlined in the test specifications, because it limits the amount of reading involved in taking vocabulary tests and therefore allows for clearer score interpretation. This advantage also means that lower level learners can be better catered for as they may lack the language knowledge to understand verbal definitions or synonyms.

Alderson (1990) further claims that computer-adaptive tests should be promoted as they provide measurements of time and therefore can yield information about fluency of access to the linguistic knowledge components. Although this is not the primary aim of the present project, designing the test in this technological environment offers the chance to integrate timed elements long-term as there may be good reasons for monitoring fluency (Segalowitz, 2015). This would not be possible if the test was laid out as a paper-pencil version. In fact, another key advantage of delivering the test in a computer-adaptive environment is that this permits a range of future improvements, such as the ability to incorporate additional word knowledge dimensions (e.g. measuring knowledge of spelling or collocations) after the completion of the doctoral project.

One of the biggest advantages of delivering this test in an online, computer-adaptive environment, however, lies in its accessibility and dissemination, according to Chapelle & Douglas (2006). A computer-adaptive test can be taken "at many convenient locations, at convenient times, and largely without human intervention" (Chapelle & Douglas, 2006, p. 23). It appears that there is also sufficient evidence to conclude that "the computer may be used to administer tests in many traditional multiple-choice test settings without any significant effect on student performance" (Paek, 2005, p. 1). Wang et al. (Wang, Jiao, Young, Brooks, & Olson, 2008) reach the same conclusion, suggesting that online test delivery should not pose any problem for learners of the current generation. Tseng (2016) thus concludes from his study that "the measurement of vocabulary knowledge has entered a new era" (p. 20), in which moving towards computer-adaptive vocabulary testing can open up new opportunities to improve our understanding of the L2 mental lexicon.

As outlined by Chapelle and Douglas (2006) the validation of a computer-adaptive test needs to address the specific concerns and validity threats related to computer-adaptive testing, which is why it is crucial to have the test set up as a computer-adaptive test from the design stages on rather than retro-engineer a finished paper-and-pencil version into an online environment. Chapelle and Douglas (2006) list several issues including the item formats, the item scoring and the algorithm for adaptive item selection that will impact on the validation of the test so that it will be necessary to design and implement these elements into the tests ab initio to provide the basis for extensive validation of a computer-adaptive test.

Tseng (2016) points out that computer-adaptive test designs can also avoid the "bandwidth-fidelity dilemma" (Weiss, 1985). A peaked conventional test design with numerous items from difficulties (or frequency bands) centering on the pre-determined level approximate to the test takers' level retains fidelity but suffers in bandwidth (more precision but within a narrow sampling area). A rectangular conventional design faces the opposite problem. It tests from a wider range, but lacks precision because fixed-length tests then only allow few items to be sampled from each ability level. Tseng (2016) argues that CAT can counter both these issues by allowing for a dynamic and flexible testing algorithm with sufficient items provided overall as well as sufficient items that are targeted at particular levels. This confirms Schmitt (2010), who argues that one of CAT's main advantages besides its adaptiveness and flexibility to enhance validity is that teachers and learners no longer have to guess the level of the test taker a priori or have to work through an entire test in a lockup fashion. Tseng (2016) states that "[c]learly, the adaptive item selection strategy taken by CAT enables a more fine-grained distinction between test taker abilities" (p. 3), which is certainly highly desirable in diagnostic testing. In contrast, fixed length formats might provide unstable and imprecise vocabulary size estimates for learner groups at either end of the ability continuum (Schultz, Whitney, & Zickar, 2014).

Despite these advantages, few computer-based vocabulary tests, let alone computer-adaptive vocabulary tests have been developed to date. Tseng's (2016) test is a notable exception, but it is designed for the national context of Taiwan and its item bank is designed accordingly. More prominently, the revised CATSS (Levitzky-Aviad, Mizrahi, & Laufer, 2014) is an internationally used computer-adaptive test. However, it is only really adaptive in terms of the modalities presented to the candidate (i.e. the "strength" dimension per item). The items themselves, however, and the progression through the frequency levels remains static. Test takers are presented with a fixed number of items for each frequency band and are all presented with the same items. It thus appears that exploring this promising but largely under-researched and underused technological advantage of computer-adaptive vocabulary testing is necessary.

6.2.The two approaches subject to comparison and their operationalisation

Operationalizing a CAT mode, however, still requires decisions about the test design. At least two approaches seem to be relevant options, which will need to be explored for the present purpose.

6.2.1. The "Floor first" approach

The first approach, or design algorithm, we will call "floor first" (FF). In this FF design, a test taker starts with a number of high-frequency items from the first band and proceeds through the bands until test takers' success rate falls below a certain percentage (as visualized in the Figure below).

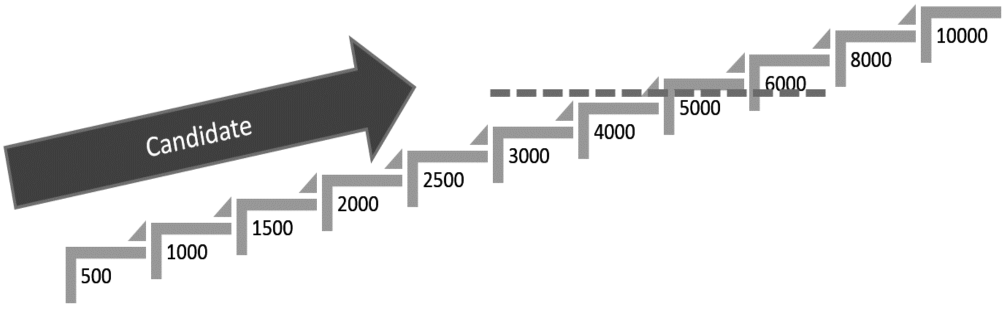


Figure 27: Schematic depiction of a FF design

In the case of the present test, this was operationalized as follows. Candidates are first presented with five items from the first band (500). If they answer all five items correctly, they then move on to the next band, in this case 1,000. In this fashion, they move quickly through the higher-frequency bands that do not pose a problem for them if they are a more proficient EFL learner. This algorithm is followed until the point where a test taker does not answer all five initial items from a band correctly. In that case, the program adapts and presents them with another set of five items from that frequency band. If their score on these second five items matches their score on the first five items (with an allowed deviation of +/-1 point), then the scores from these two rounds are added together and the sum is recorded as their score for that level (in percentage form, adjusting for the increased number of items answered vis-à-vis the previous levels). They then move on to the next level, where the algorithm applies the same rules. If their scores from the two rounds in a frequency band deviate by more than 1 point, they are presented with a third set of five items from that frequency band. Regardless of their score on these third five items, their total score out of the now 15 items is recorded as band (percentage) score and they move on to the next frequency band with the algorithm applying the same rules again. The test terminates if a test taker scores below 20% correct on a total of 15 items from a band. Candidates are, however, given a third set of five items if they scores 20% or below in total in the first two sets. The table below exemplifies a possible test progress of a test taker.

Table 32: Potential test taker progression in FF design

Frequency band	Round 1 (1^{st} 5)	Round 2 (2^{nd} 5)	Round 3 (3^{rd} 5)
1	5/5		
2	5/5		
3	4/5	4/5	
4	3/5	5/5	3/5
5	4/5	1/5	2/5
6	1/5	1/5	3/5
7	1/5	0/5	1/5
8			

In case a very proficient test taker manages to answer all items correctly in all bands, they will be presented with another 10 items each from the lowest two frequency bands. This means that test length and the number of presented items is adaptive and will vary with the proficiency of a candidate. At a minimum, a candidate will be presented with 75 items. At most, the test will administer 165 to any one candidate. However, particularly the latter scenario is rather unlikely as it would be unusual for a test taker to already struggle at the level of 500 but still be able to stay above the 20% accuracy threshold until the lowest frequency band at 10,000 lemmas.

Admittedly, selecting five items per round per level is an arbitrary decision that can be altered and probed further once a beta version of the test is running and has sufficient validation evidence behind it. It may be that four items per round could be enough, or it may be that six or more items per round yield better results, psychometrically speaking, while still being doable for test takers within a reasonable time. The current decision to trial with five items per round was informed by a) the fact that Gyllstad, Vilkaité and Schmitt (2015) recommend 30 items per 1,000 level as good coverage, which would be reached for the high-frequency levels if all three rounds per band were administered, and b) an awareness of the trade-off between total testing time and amount of items administered. While it seems important to gather as much information about the vocabulary knowledge as possible, it is also key to keep the total testing time under 30 minutes to minimize fatigue or demotivation, even if the maximum number of items were to be administered.

6.2.2. The "multi-stage multi-level" approach

By contrast, a multi-stage multi-level design (MSML) (Luecht, Brumfield, & Breithaupt, 2006; Luecht & Nungester, 1998; Zenisky, Hambleton, & Luecht, 2010) provides the candidates with items from a range of bands in a first stage and then proceeds to further stages with the range of items getting more narrow in the process.

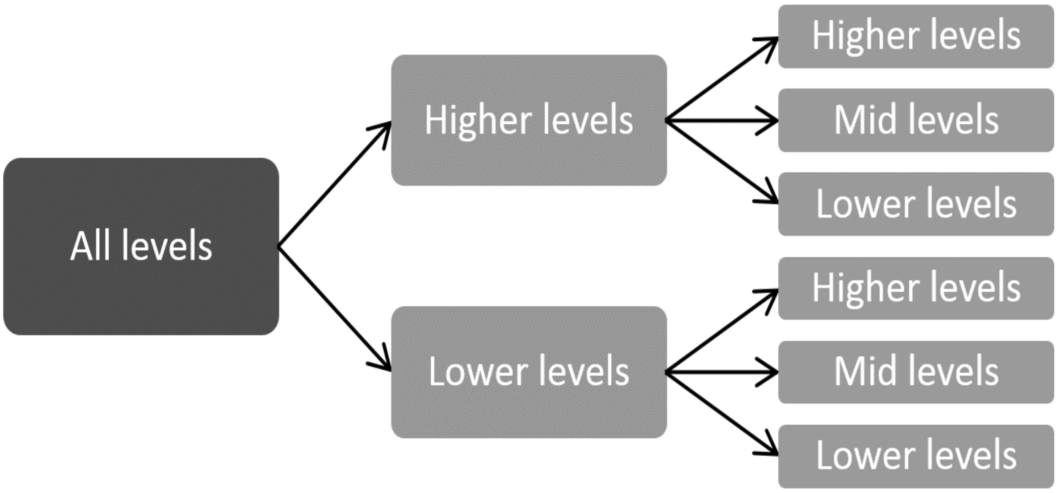


Figure 28: Schematic depiction of a multi-stage, multi-level design

For the purpose of the present study, this design was operationalized as follows. In a first stage, candidates are presented with five items from each of all the frequency levels. Their score on these 55 initial items determines their "base" frequency level for the second stage. Every five correct answers thereby represent one frequency level. For instance, if a test taker scored 32 out of the 55 correct, their "base" frequency level would be determined as band 7. In the second stage, test takers are then presented with another five items from their "base" frequency level, plus another five items each from the two adjacent levels below their "base" and the two adjacent levels above their "base". In the example of a test taker with 32 points in the first stage, these would be five items each from the levels 5, 6, 7, 8, and 9. The test taker's score on these 25 items are treated similarly to the scores from stage one. Their total score again determines the "base" band for the third stage. If the example test taker scored 14 out of the 25 items in stage two, then their "base" would be determined as band 7. In the third stage, candidates are then presented with another five items from their "base" level from stage 2, and another five items each from the two adjacent frequency bands (one above, one below). The test terminates after these additional 15 items. Test length is therefore fixed at a total of 95 items, irrespective of candidate ability. Only the focus of the items changes with test taker proficiency to home in on some frequency bands that might be of particular interest. Table 33 illustrates a possible test taker progression through this design.

Table 33: Potential test taker progression in MSML design

Frequency band	**Stage 1 (55)**	**Stage 2 (25)**	**Stage 3 (15)**
1	5/5		
2	5/5		
3	4/5		
4	4/5		
5	4/5	4/5	
6	3/5	3/5	4/5
7	3/5	4/5	2/5
8	2/5	1/5	1/5
9	1/5	2/5	
10	1/5		
11	0/5		

While there are theoretical advantages and drawbacks to either of these approaches, the aim of the current study was to establish empirically which of these two algorithms would produce the more useful measurements. For the purpose of this comparison, more useful was defined as producing a) more reliable results over two administrations, and b) more representative results of a larger item pool. The research questions were thus formulated as follows:

> RQ1) Which of the two test designs (FF or MSML) has better test-retest-reliability at the individual frequency levels?
>
> RQ2) Which of the two test designs (FF or MSML) produces scores that are more representative of a larger number of items at each frequency level?

6.3.Methodology

To investigate RQ1 the test in the two versions was administered twice to an EFL learner population. After cleaning the data, 85 EFL learners from three different L1 backgrounds (German, Hungarian, and Arabic) remained that had taken the FF version twice immediately after each other. Most of these candidates (79%) were high-school students nearing the end of their secondary education. The other 21% were BA students of English at a Saudi Arabian university. The group's mean age was 16.99 years (SD=1.28) and they had been learning English for 7.84 years on average (SD=1.65). 52% of the participants in this group were female, 48% were male. 72 EFL learners took the MSML version twice immediately after each other. They came from three different L1 backgrounds (Croatian, German, and Arabic). About half of these EFL learners (55%) were high-school students nearing the end of their secondary education. The other half were BA students of English. The group's mean age was 19.25 (SD=3.01). They had been learning English for around 9 years (M=9.29, SD=2.85). 77% of the test takers in this group were female, 23% were male. Students' scores were only linked through an ID code, which was assigned by the invigilators and unknown to the researcher. The candidates' scores from the two attempts were then correlated to establish which version showed better retest-reliability.

To probe RQ2, 34 different EFL learners from Lithuania (2 L1s: Lithuanian and Polish) participated in a separate study. These candidates were all BA students of English at a Lithuanian university. Their mean age was 19.66 (SD=1.72) and they had been learning English for 11.4 years on average (SD=1.96). 91% of these participants were female, 9% male. These students first took the regular FF version and then the regular MSML version. They were then instructed to take a third version of the test, which was programmed so that depending on how many items they had answered in the first two versions, they would be presented with additional items for each level until they had answered 25 items per level. For example, if a candidate had only answered five items from the 500 band in the FF version (because they answered them all correctly), and had only answered five items correctly in the MSML version from the 500 band (because they were assigned a higher "base" level after the first stage), then they were presented with another 15 items from the 500 band in version three to make up a total of 25 items per band. If, however, they had already answered 15 items in the FF version at the 6,000 band and were presented with at least 10 items from the 6,000 band in the MSML version, then they were not presented with additional items from that level when taking version three. Afterwards, the band percentage scores of these participants on each of the first two versions (FF and MSML) were correlated with the scores on the 25 items per frequency band to investigate which version better represented a larger item pool and thus would produce more robust measurements. This approach seemed favourable to determine the validity of the band scores than, for instance, correlating the band scores with band scores on other external criterion measures, such as the VLT or the VST, as they both differed in their construct too considerably to offer a valuable reference point. In addition, the functioning of the items in the item pool from which these larger band sets were sampled had already been established in the piloting phase. However, a serious limitation besides the relatively small sample size needs to be mentioned at this point. Due to the size of the item pool it is probable that some items would have been administered more than once per candidate in this research design. For reasons of funding, however, it was not possible to implement this restriction in the item sampling for the research design. Nevertheless, it is highly unlikely that a test taker would have encountered exactly the same 15 items in each of the test forms, which is why the total score on the 25 can arguably still be taken as a relatively solid indication of vocabulary knowledge of the broader frequency band. Therefore, the results are valuable for the purpose of pointing towards which of the two algorithms should be selected for the final test construction. However, follow-up studies with an even larger item pool certainly need to corroborate this further for the selected algorithm.

6.4.Results

Table 34 displays the results of the two attempts of the 85 test takers in the FF version. The mean % scores per band for the two attempts show a relatively clear decline of scores across the frequency bands in the first attempts with the highest % scores in the high-frequency bands and a steady decrease in mean % score per frequency band. The pattern emerges less clearly in the scores of the second attempt with a slight unexpected spike of mean % scores at bands five, seven, and nine. While mean scores were very high at the first three levels, i.e. the 500, 1,000 and 1,500 lemma band, participants still scored 51.70%, and 51.82% respectively, correct at the lowest frequency band of 10,000 lemmas. Percentage scores were higher than 90.71% in both attempts for the three high-frequency bands. The correlations between the two attempts were significant at all frequency levels. The coefficients ranged from .38 (band 1) to .92 (band 10), with most coefficients being above .87. The coefficients are relatively low in the high-frequency bands, particularly band 1, because of the high proficiency of the participant group. Since they answered most items at this level correctly, there was limited variance, which is why the low correlation coefficient of .38 is hardly surprising.

Table 34: Mean % scores and correlations between attempts per band (FF version)

Frequency band	Mean % (SDs) [1]	Mean % (SDs) [2]	Correlation
1	95.63 (8.0)	95.33 (7.8)	.38**
2	90.12 (17.2)	92.88 (14.0)	.56**
3	90.71 (17.6)	90.94 (19.1)	.71**
4	77.69 (26.3)	76.71 (27.6)	.79**
5	77.57 (32.5)	77.61 (32.4)	.91**
6	69.18 (35.5)	68.51 (35.0)	.91**
7	67.88 (36.8)	70.00 (36.7)	.87**
8	64.71 (36.2)	60.24 (35.0)	.88**
9	63.88 (38.3)	61.22 (38.5)	.90**
10	57.00 (34.0)	54.80 (35.9)	.92**
11	51.70 (32.5)	51.82 (34.0)	.90**

The MSML results of the other 72 EFL learners are less clearly interpretable. There appears to be a general trend of declining mean percentage scores as participants move towards the lower frequency bands. However, this is much less linear. While the mean % at band 2 is 94.72%, and 94.42% respectively, the mean score, somewhat unexpectedly, increases again at band 3 to 96.81%, and 98.26% respectively, thereby even surpassing the mean scores of band 1 (96.67% and 95.37%). After this, the mean % scores follow a more predictable and expected pattern. However, the correlations between the two attempts are not significant across all of the frequency bands. For bands 2 and 3 they fail to reach significance. In addition, the correlation coefficients are considerably lower than for the FF versions. Only at three bands the coefficients are higher than .70. Interestingly, however, there is a remarkable difference in mean % scores between the FF and the MSML version at the low frequency end. While the mean % score drops below 70% in the FF version already at band six, the 72 participants in the MSML version do not drop below this threshold even at band 11. While it is certainly possible that the MSML group was more proficient than the group that took the FF version, this cannot be ascertained from the study design as there was no overlap of candidates taking both versions twice. It seems, however, that the fact that learners are presented with items from all frequency bands in the initial stage of the MSML design could allow learners to show more of their knowledge than the FF design, which terminates the test before the low levels would be presented to them.

Table 35: Mean % scores and correlations between attempts per band (MSML version)

Frequency band	Mean % (SDs) [1]	Mean % (SDs) [2]	Correlation
1	96.67 (8.2)	95.37 (10.0)	.36**
2	94.72 (12.1)	94.42 (10.3)	.21
3	96.81 (8.7)	98.26 (5.8)	.22
4	87.80 (17.5)	89.72 (16.4)	.39**
5	90.14 (15.2)	89.72 (16.0)	.40**
6	87.03 (16.1)	85.42 (17.9)	.61**
7	86.01 (14.6)	85.83 (15.0)	.27*
8	77.88 (19.3)	80.92 (19.3)	.49**
9	78.28 (19.8)	79.48 (20.0)	.79**
10	70.83 (17.7)	73.91 (18.8)	.77**
11	70.73 (17.9)	70.68 (19.9)	.70**

To answer RQ2, the scores of 34 candidates on both the FF and the MSML version were compared with the scores on a larger item sample of 25 items per frequency band. The scores of these 34 learners seems to suggest that the MSML group was not more proficient than the FF group discussed above as the level mean % scores are very similar. Had it really been the case that one design was allowing for higher scores than the other, this would show in these scores as these are based on identical candidates taking both versions. Also, the purpose of RQ2 in this design was to guard exactly against this by checking which design's scores provide a better representation of the candidates' actual band knowledge (as indicated by the score on a larger item sample of 25). The progression, or rather the expected decline, in mean % scores is again not following a very clear pattern. Scores seem to spike slightly at bands three, five, and most surprisingly band eleven. The correlations between the FF version and the 25 items per band are all significant and range from .53 (band 5) to .97 (band 11). The majority of correlation coefficients are above .81. The correlations between the MSML scores and the scores on the 25 items per band are less straightforward. At band 3, the correlation fails to reach significance with a coefficient of only .12. All other band score correlations are significant, but in seven out of ten cases, the coefficients are lower than for the FF-25 correlations. Only for bands eight, nine and ten, are the correlation coefficients higher in the MSML-25 correlation than in the FF-25 correlation. This, again, could be a product of the proficiency of the candidate group, as there were more items sampled in these bands.

Table 36: Mean % scores for FF, MSML, 25-item-version and correlations

Freq. band	FF	MSML	25	Correlation FF-25	Correlation MSML-25
1	96.9 (6.0)	97.1 (8.7)	96.3 (4.5)	.77**	.65**
2	95.3 (8.3)	93.5 (12.8)	94.9 (6.0)	.73**	.73**
3	98.5 (6.1)	99.4 (3.4)	98.4 (3.4)	.81**	.12
4	91.2 (14.5)	91.2 (14.9)	90.7 (11.3)	.91**	.82**
5	95.6 (7.5)	96.5 (10.4)	94.2 (7.1)	.53**	.53**
6	88.8 (16.5)	87.2 (18.1)	89.1 (11.7)	.92**	.78**
7	88.2 (16.4)	89.6 (14.7)	86.0 (13.1)	.87**	.71**
8	85.3 (18.1)	86.7 (16.6)	83.7 (13.7)	.79**	.89**
9	83.5 (18.9)	84.7 (17.9)	83.5 (14.7)	.81**	.88**
10	77.6 (16.5)	75.0 (18.7)	76.4 (15.9)	.91**	.94**
11	81.5 (15.7)	84.6 (13.4)	82.8 (14.0)	.97**	.81**

6.5.Discussion

The results appear to suggest that the FF design is a more useful approach for developing the diagnostic computer-adaptive vocabulary test. The FF algorithm consistently outperformed the MSML algorithm in terms of test-retest reliability. The correlation between the two test administrations' scores was higher across all frequency bands, thus indicating that the FF version produces more consistent measurements. In the MSML version's two administrations, some frequency band % scores did not even reach significance. In terms of representativeness, the FF also seemed to fare better than the MSML version. Although not as clearly superior to the MSML as in the reliability investigation, the FF scores did, overall, correlate more strongly with the overall frequency band scores from the larger item pool. In 8 out of 11 bands, the FF scores corresponded more closely to the frequency band scores gained from a larger item sample. This may indicate a potentially higher validity of the FF scores.

It must be noted here, though, that there may be a case for different algorithms being a better fit for different populations. It might be that a MSML design is to be favoured for particular proficiency groups, or even for particular L1 groups. Cobb (personal communication) points out that, for instance, French-speaking learners of English might benefit from a MSML approach more than German-speaking learners of English do. French L1 speakers may know more words that are cognates from Romance languages, which tend to appear in lower-frequency bands. In an FF design, these learners might not be able to demonstrate their knowledge of these lower-frequency words in English if the test terminates before they arrive at the relevant frequency levels. While this seems intuitively plausible, this hypothesis would need to be investigated empirically with different L1 groups. The current research design and candidate recruitment did not allow for looking into an L1 group bias in the different design algorithms.

In summary, however, the FF design appears to have key advantages over the MSML design for the present purpose. This is despite the fact that the MSML does allow to capture a broader range of items and levels by default and potentially allows homing in on a few particularly relevant levels for the learner. The MSML design thereby enables learners to demonstrate knowledge even limited knowledge at even lower frequency levels, which is likely to be mirroring the progress of vocabulary acquisition. However, being presented with many words that are far beyond one's ability bears also a great risk of demotivating learners. In the FF design, the learner is presented with words that are challenging, but the test stops before it may become too discouraging (i.e. when the learner falls below a 20% accuracy threshold). Given the established importance of motivation in language learning (Dörnyei, 2014; Dörnyei & Ushioda, 2011), this avoidance of negative impact appears key to take into account when selecting the algorithm. Not presenting learners with tasks or items that are highly likely to be unachievable should therefore have a positive influence on their L2 motivational self system (Dörnyei, 2009), particularly their L2 learning experience as motivation is more likely to be maintained and protected in this way. Also, while the MSML design has a fixed test length, the FF design is potentially shorter in length. This practical aspect also needs to be considered. Above all, however, the present data has borne out that the FF design generates more reliable and also generally more representative, and thus valid scores.

The aim of the study presented in this chapter was to arrive at a first indication of which design algorithm is more promising for the present test development project. It did by no means attempt to arrive at a conclusive answer of which design is generally the best for vocabulary tests. Despite the limitations in research design outlined above and the relatively small sample size of candidates, it did provide a number of very useful insights and pointers as to how to move ahead with the test design. It did, however, also indicate that there will be a persisting need to corroborate both the general item quality as well as the usefulness of the FF algorithm with learners from a broader range of proficiencies. Particularly lower-level test takers will have to confirm these initial decisions before the test can be officially launched.

Because of this relatively homogenous proficiency group, the present study was also unable to answer the question whether fixed test length is an advantage or disadvantage. It remains to be investigated what the average test length and duration will be when administered to very low-level and very proficient learners. In addition, it is noteworthy that neither of the designs examined here was making full use of the potential of the IRT information available from the pilot and further administrations. Usually, this is a defining feature of computer-adaptive tests. The logit difficulty information is used by the system to find exactly the items that match the person's ability depending on their previous responses and thus generate a test that will yield the maximum of psychometrically useful information for the end user. Tseng's (2016) computer-adaptive vocabulary test, for instance, does this using Bayesian statistics to shorten test length. The design of the present test, however, is predominantly frequency based and clusters items in sets per frequency bands. While the computer-adaptive version of the Word Parts Levels Test (Mizumoto, Sasao, & Webb, 2017) has very recently showed how a levels-based approach might be married with the use of IRT data, this was beyond the remit of this study due to the size of the test taker sample. It therefore remains to be explored how to integrate the IRT information in the test administration once the psychometric properties of the item pool can be considered stable enough through more extensive administration.

6.6.Summary

The purpose of the study presented in this chapter was to compare two different computer-adaptive algorithms for implementation in the test design: "floor first" versus multi-stage multi-level. Two separate studies found that a "floor first" design, in which candidates progress from high-frequency items to lower-frequency items until they fall below a predetermined accuracy threshold, yielded both more consistent and more representative scores. The "floor first" algorithm was established to have good test-retest reliability across all frequency levels. The findings from the smaller-scale study also showed that this design gives a good estimation of test takers' vocabulary knowledge at each of the frequency levels, as indicated by significant and high correlations with scores from a larger item sample from each frequency band. It was thus decided to move forward implementing this "floor first" design in the diagnostic computer-adaptive vocabulary test on the basis of the findings presented in this chapter as well as other practical and, more importantly, motivational reasons. Based on this final design decision, the test specifications were therefore adapted to the following:

Table 37: Revised test specifications

Diagnostic vocabulary test - Specifications	
General purpose	To diagnose the written receptive lexical abilities of EFL learners
Specific purpose	• To determine whether EFL learners know the written form-meaning link to the extent that it would allow employing that vocabulary knowledge for reading comprehension • To determine how well EFL learners know the form-meaning link of words from different frequency levels up to the first 10,000 content lemmas
Target language situation	International learners of EFL
Description of the test taker	All ages, but likely to be age 10 and upwards; international audience, diverse L1 backgrounds, beginner to (upper-)intermediate proficiency level
Test source	Discrete items sampled from the first 10,000 content lemmas (nouns, verbs, adjectives, adverbs) of the lemmatized COCA word frequency list (Davies, 2008-) Items will be clustered and sampled from the following frequency bands 1. 1-500 2. 501-1,000 3. 1,001-1,500 4. 1,501-2,000 5. 2,001-2,500 6. 2,501-3,000 7. 3,001-4,000 8. 4,001-5,000 9. 5,001-6,000 10. 6,001-8,000 11. 8,001-10,000

Item format	Four-option multiple choice (three distracters), target item presented in short, non-defining context, distracters either picture-based (in the first 1,500 lemmas) or text-based (synonyms, definitions) Distractors will be based on lemmas from the same frequency band that are plausible within the context of the example sentence but unrelated to the meaning of the target. Whenever an item cannot be defined in words that are of a higher frequency than the target, a picture must be used
Items per level	5-15, dependent on candidate answers Computer-adaptive algorithm in "floor first design": Progression from high-frequency to lower-frequency bands 5 items presented first - if all correct, then move to next level - if not all correct, then 5 more items from same level - if score on 1st 5 = score on 2nd 5 (+/-1), then move to next level - if score on 1st 5 ≠ score on 2nd 5 (+/-1), then 5 more items from same level - after max. 15 items per level record % score and move to next level If all items in test correct, then present another 10 items per level for two lowest frequency
Total number of items	Computer-adaptive, dependent on candidate answers Minimum: 75 Maximum: 165
Instructions	Target language with one example
Weighting	1 point per item
Time allowed	Untimed, but probably no longer than 30 minutes in total
Administration	Computer-delivered and scored through online website
Score reporting	In diagnostic frequency profile, split into bands and linked to information about lexical requirements of different communicative abilities (coverage research) and CEFR levels

This test form was now ready to be validated, the first step of which will be discussed in the following chapter.

7. Relationship of vocabulary scores and reading test scores

This chapter explores the link between the scores of the present vocabulary test and a CEFR-based reading comprehension test as part of an initial validation effort. First, the research related to the relationship between vocabulary knowledge and reading comprehension will be reviewed and the importance of linking vocabulary test scores to reading comprehension in validation will be established. It will then present a study that investigated the vocabulary knowledge profiles of readers at different CEFR levels according to a standardized proficiency test. It will probe whether the vocabulary test successfully distinguishes between the vocabulary knowledge profiles of readers at different levels so as to provide backing for the validity of the newly developed instrument.

7.1.Research on the relationship of vocabulary knowledge and reading comprehension

"Nobody interprets the [vocabulary test] scores as simply words that learners can answer on a vocabulary test" (Schmitt, 2014, p. 943). Instead, vocabulary test scores are often interpreted to inform the score user about the test taker's ability to employ their vocabulary knowledge in a language skill or in performing a linguistic task. It is pivotal therefore for a test of written receptive vocabulary to demonstrate how the vocabulary test scores actually link to reading comprehension scores.

Although the present test is not conceptualized as a test of vocabulary size, its construct could be equated with that of vocabulary breadth as it only assesses the mastery of form-meaning link knowledge per frequency band. Vocabulary breadth, or knowledge of the form-meaning link of many words, has been demonstrated to be integral to successful use of any language skill (e.g. Alderson, 2005; Daller et al., 2007; Meara, 1996; Nation & Webb, 2011; Schmitt, 2010). Particularly, the relationship between vocabulary knowledge and reading comprehension has been researched extensively.

Anderson and Freebody (1983) state that "people who do not know the meaning of very many words, are most probably poor readers" (p. 367). It can be argued that “one’s level of vocabulary is highly predictive, if not deterministic, of one’s level of reading comprehension” (Sternberg, 1987, p. 90). Koda (2004) maintains that “text-meaning construction is virtually impossible without functional knowledge of the words appearing in the text” (p. 256) and that “successful comprehension is heavily dependent on knowledge of individual word meanings” (p. 48). Several studies have supported these claims. Henning (1975) showed through regression analyses that L2 vocabulary knowledge is the key predictor of reading comprehension at intermediate level. Pike (1979), Barnett (1986), and Koda (1989) also found that L2 vocabulary knowledge contributed significantly to L2 reading comprehension. Parry (1987) maintains that unknown word meanings will lead to misinterpretations of whole texts, thus supporting the key role of vocabulary knowledge. Coady et al. (1993) also demonstrated that vocabulary gains have a significant positive effect on reading comprehension. Alderson (2000) reports that “factor analytic studies of reading have consistently found a word knowledge factor on which vocabulary tests load highly” (p. 99). Vocabulary knowledge can therefore be seen as an important predictor of variance in reading test performances (Qian, 2002). Indeed, Schoonen et al.’s (Schoonen, Hulstijn, & Bossers, 1998) findings highlight that L2 vocabulary knowledge is a pivotal predictor of L2 reading. Nassaji (2003) even goes so far as to say that vocabulary knowledge is the strongest predictor of the component skills associated with reading in an L2.

Yamashita (1999) concludes from her analysis of Japanese English as a foreign language (EFL) readers that L2 vocabulary knowledge explains L2 reading comprehension score variance to a large extent, particularly for higher ability readers. Brisbois (1995) reports similar findings in her investigation of English-speaking learners of French, as do Van Gelderen et al. (2003, 2004). Kremmel, Brunfaut and Alderson (2015) also found that L2 vocabulary knowledge, as tested by the DIALANG vocabulary test, emerged as a crucial component of reading ability in their structural equation model. A study by Laufer (1992b) indicated strong correlations between L2 vocabulary size and L2 reading comprehension, which led her to claim that the hypothesized linguistic threshold for L2 readers to pass before being able to transfer their L1 reading strategies is mainly a lexical threshold (Laufer, 1992a). Jeon and Yamashita (2014) echo this with the finding that L2 vocabulary knowledge was one of the three strongest correlates of L2 reading comprehension in their meta-analysis.

While the relationship of the test scores of this newly developed test are not only interesting in terms of being able to link scores at the respective levels to coverage research (Schmitt, Gardner, & Davies, under review) and therefore facilitated score interpretation in terms of what learners can use the vocabulary knowledge for, it seems also a relevant part of a validity argument that the test would be able to discriminate between readers at different levels. Huhta, Alderson, Nieminen and Ullakonoja (2011) have already demonstrated that this might not be as straightforward as one might hope. Despite the average scores following the expected pattern across frequency bands and proficiency levels, their adapted L2-L1 version of the VLT only managed to significantly distinguish between some of the proficiency levels at a few of the band scores. However, this may only support the notion that not all in reading comprehension can be accounted for merely by vocabulary knowledge. Particularly so, because they did find a consistent, linear relationship between vocabulary scores and reading comprehension across bands and reading ability levels. Whether such a relationship can be found also for the present test is the focus of the study outlined in this chapter. Specifically, it aims to answer the following two research questions:

1. What is the relationship between reading ability (in terms of CEFR level) and vocabulary knowledge scores at different frequency bands?
2. What are the typical vocabulary knowledge profiles of readers at different CEFR levels?

7.2.Methodology

To investigate these two research questions, the newly developed vocabulary test was administered alongside a standardized reading test. Both were administered online, first the reading test and then the vocabulary test. Candidate scores were linked through a unique ID code provided by the test provider of the reading test. The vocabulary test was administered in the design outlined in the summary of the previous chapter. The Aptis reading test was used as a measure of reading comprehension. As already described in detail in Chapter 4.5, the Aptis reading test is a multilevel test that is part of a test suite designed by the British Council (O'Sullivan, 2015) for listening, reading, speaking, and writing. It intends to measure reading ability from A1 to the C1 level on the Common European Framework of Reference (Council of Europe, 2001). Scores are reported both as numerical values on a scale from 0–50 and as a CEFR level from A1 to C. Candidates were asked to complete all four tasks of a reading testlet, which is usually administered alongside a "CORE" grammar and vocabulary test. The "CORE" component, however, was deactivated for the purpose of this investigation and candidates only took the reading component. A detailed description of the tasks can be found in Section 4.5 of this thesis and in the Aptis technical manual (O'Sullivan & Dunlea, 2015).

Participants for this study were recruited from an intact Austrian high-school class and from an Iranian university's medical studies program. The 15 Austrian EFL learners (7 female, 6 male, 2 did not indicate) were in their penultimate year of secondary schooling. They had a mean age of 16.3 years (SD=.6) and had been learning English, on average, for 7.4 years (SD=1.3). Most of them (86.7%) were German native speakers, one indicated Spanish as their L1 and one Dutch. The 68 medical students from the Iranian university (41 female, 27 male) had a mean age of 21.6 years (SD=3.0) and had been learning English, on average, for 7.5 years (SD=4.2). Most of them (97.1%) were Persian L1 speakers. One student indicated Bengali as their L1 and one student Arabic. It was hoped that by recruiting from these two very different groups, a range of proficiency levels would be covered.

Reading score results and vocabulary test results were analysed with SPSS® 22. Given the non-normal distribution of scores per level and the small sample sizes in each proficiency group, a non-parametric Kruskal-Wallis test with post-hoc Mann-Whitney-U tests between group pairs was used to examine whether there are clear distinctions between the vocabulary score of candidates at different CEFR levels. A Bonferroni adjustment was applied to control for Type I errors in multiple comparisons.

7.3.Results

The reading test scores showed that most candidates were at a CEFR B level. In the Austrian group, 40% of participants scored a B1, and 40% were attributed B2 level in reading. The reading scores of the Iranian group were slightly better, with 27.9% scoring a B1, 42.6% scoring a B2 and even 14.7% reaching a C level in the Aptis reading comprehension test. In total, this means, that of the 83 test takers in this study, 2 were classified at A1 level, 11 candidates emerged as being A2 level readers, 25 as B1 level readers, 35 were B2 level readers, and 10 scored a C level in the Aptis reading comprehension test. While a more balanced distribution of proficiencies would have been desirable, there is still some tentative validation evidence to be gleaned from the following analyses.

In a first step, the average % scores per frequency band were plotted for each reading CEFR level (Figure 29). This was done to arrive at an overview of how the test performs with learners at different proficiency levels. The data for this is presented in Table 38 below.

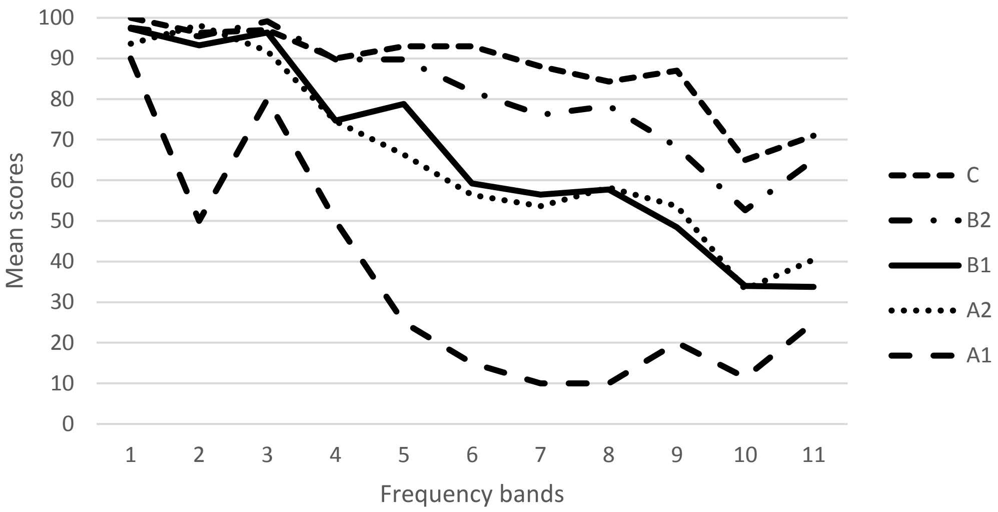


Figure 29: Mean % scores per frequency band of readers at different CEFR levels

Figure 29 shows that the test mostly performs as expected. In general, readers at C level show a better vocabulary knowledge than readers at B2 level. B2 readers, in turn, show better vocabulary knowledge than B1 and A2 readers. The two A1 readers also perform the weakest in all frequency bands on the vocabulary test. However, the sample size for this group is so small that it can be disregarded, even though the profiles of these two individual learners do spike at unexpected points. Nevertheless, there are some noteworthy general observations.

As can also be seen in Table 38, the frequency effect only seems to come into play after the 1,500 band. Readers at the proficiency levels measured here seem to perform very similarly at the highest frequency bands. At band 2, A2 readers even seem to outperform C level readers, which is surprising, but taking the standard deviations into account, there is little difference at the first three frequency bands between readers from A2 through to C level. This warrants closer inspection with a larger population sample per CEFR group. The curves of C readers and B2 readers only seem to diverge as of the fifth band (2,500), while the vocabulary profiles of A2 and B1 readers appear very similar. However, there are three bands, in which A2 readers unexpectedly fare better than B1 readers: bands 2, 9, and 11. Again, a larger sample would be necessary to corroborate this. Second, there appears a spike in mean % scores across all CEFR levels except B1 for the final band 11. It would be expected that the average score on this lowest frequency band is the lowest overall. This highlights the need for closer inspection of the items at this level with a larger population.

The same data can also be presented differently in graphical form for additional useful information about the test's validity and the relationship of scores to reading levels. When looking at the data per frequency band and the expected increase in mean % scores across the five tested CEFR levels, one can see that for most bands there is a clear progression from the mean % scores of A level readers to C level readers. Figure 30 illustrates this. Even though these progressions are not as neatly linear as those found by Huhta et al. (2011), they are generally behaving as expected and thus tentatively confirming the functioning of the vocabulary test. The band 1,000 score for the A2 readers is an anomaly in this respect, but given the small size of the A2 reader group, and the SD associated with the scores of most proficiency groups at this frequency band, this alone does not appear cause for too much alarm. What is certainly more concerning is that at most frequency bands (2,000, 3,000, 4,000, 5,000, 6,000, 8,000, and 10,000), the test does not seem to discriminate very well in terms of vocabulary breadth between A2 and B1 readers. The mean % score increases are evident but minimal. At the bands 5,000, 6,000, and 10,000, however, the B1 readers surprisingly performed worse than the A2 readers.

Table 38: Progression of mean % scores per level across reading CEFR levels

	A1		**A2**		**B1**		**B2**		**C**	
	M	SD	M	SD	M	SD	M	SD	M	SD
500	90.0	(14.1)	93.6	(6.7)	97.3	(7.3)	97.5	(4.9)	100	(0.0)
1,000	50.0	(14.1)	98.2	(4.0)	93.2	(10.3)	95.4	(7.0)	96.3	(6.7)
1,500	80.0	(14.1)	91.8	(10.8)	96.4	(9.5)	99.1	(2.8)	97.0	(6.7)
2,000	50.0	(0.0)	74.5	(19.7)	74.7	(19.4)	89.7	(15.4)	90.0	(16.3)
2,500	25.0	(7.1)	66.4	(22.5)	78.8	(17.6)	89.7	(16.0)	93.0	(14.9)
3,000	15.0	(7.1)	56.4	(25.0)	59.2	(21.8)	81.9	(22.1)	93.0	(14.9)
4,000	10.0	(14.1)	53.6	(27.3)	56.5	(24.5)	76.1	(23.4)	88.0	(14.0)
5,000	10.0	(14.1)	58.2	(19.9)	57.7	(29.9)	78.3	(20.4)	84.3	(20.6)
6,000	20.0	(28.3)	53.6	(26.9)	48.4	(26.9)	68.2	(25.7)	87.0	(15.7)
8,000	11.4	(16.1)	33.2	(24.6)	34.0	(23.2)	52.6	(21.0)	65.0	(20.2)
10,000	25.0	(35.4)	40.5	(26.3)	33.8	(26.5)	65.0	(22.8)	71.0	(12.9)

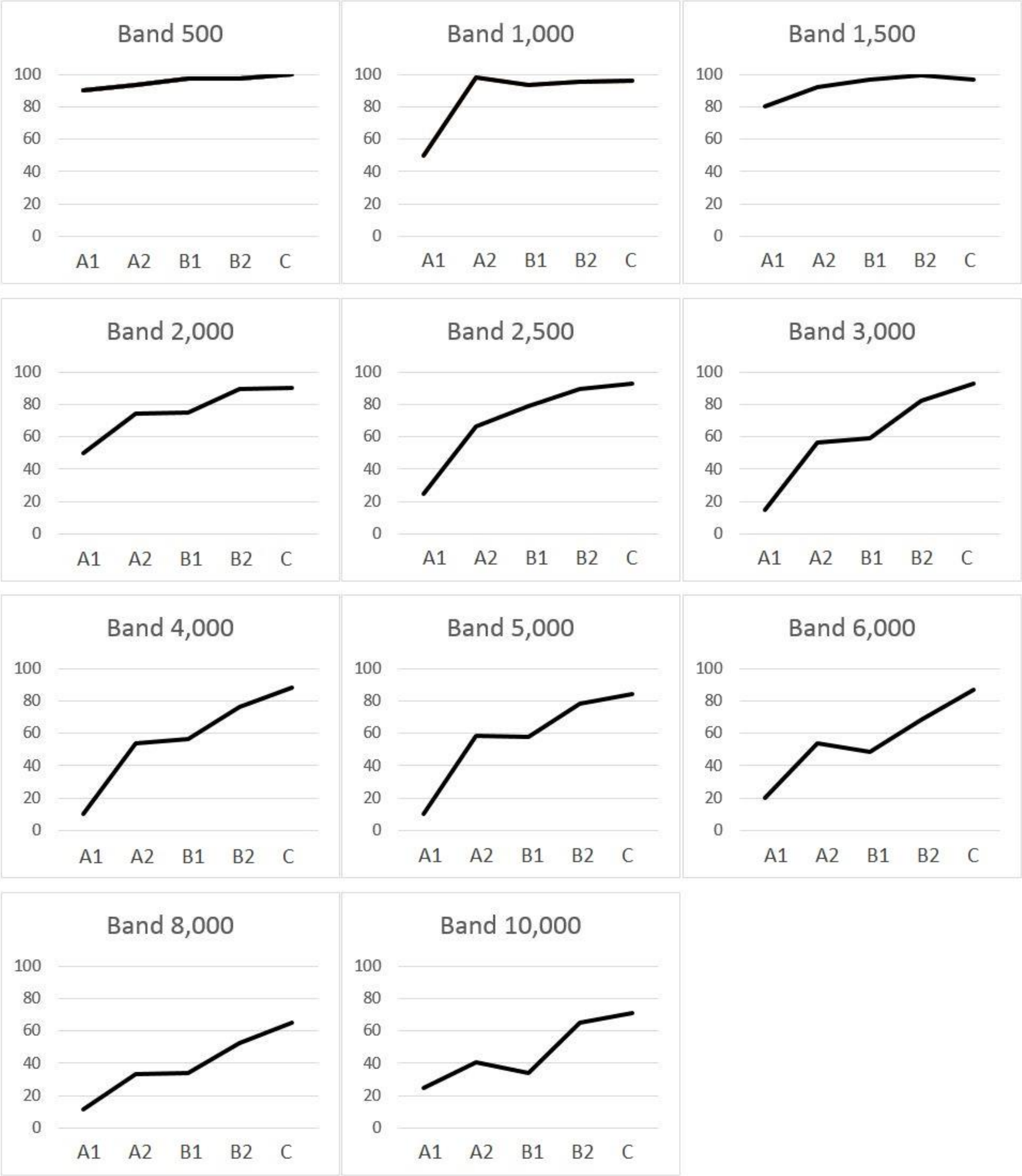


Figure 30: Progression of mean % scores per level across reading CEFR levels

For the Kruskal Wallis test, the A1 group was excluded as it only consisted of two learners. The test revealed that there was a significant difference in mean scores across the CEFR levels in all frequency bands except band 2 (1,000). The χ^2 values and significance levels are summarized in Table 39.

Table 39: Kruskal Wallis test for differences in mean scores across four CEFR levels

Band	χ^2	df	Sig.
500	9.834	3	.020
1,000	2.743	3	.433
1,500	8.494	3	.037
2,000	13.665	3	.003
2,500	18.853	3	<.001
3,000	24.745	3	<.001
4,000	17.756	3	<.001
5,000	13.620	3	.003
6,000	16.372	3	.001
8,000	18.505	3	<.001
10,000	27.109	3	<.001

However, the post hoc comparisons between pairs of proficiency levels showed that these differences only held for some CEFR level comparisons. When comparing A2 and B1 level readers, a significant difference could only be found for band 500 (p=.03). However, this was above the Bonferroni adjusted significance level at p=.008 (six comparisons). All other bands were non-significant, even at the much less conservative p-level of .05. The comparison of the A2 and B2 groups yielded (Bonferroni adjusted) significant differences at bands 1,500 (p=.003), 2,500 (p=.001), 3,000 (p=.004), 8,000 (p.004) and 10,000 (p=.002). Unsurprisingly, in all of these frequency bands, the B2 group significantly outperformed the A2 group. The findings for the A2-C comparison were similar. Significant differences between the two groups were found at bands 500 (p=.007), 2,500 (p=.004), 3,000 (p=.002), 4,000 (p=.005), 6,000 (p=.007), 8,000 (p=.004), and 10,000 (p=.003), again all in favour of the C level group. The test scores were also significantly different for the B1 and B2 groups at the bands 2,000 (p=.002), 2,500 (p=.006), 3,000 (p<.001), 4,000 (p=.003), 5,000 (p=.007), 8,000 (p=.005), and 10,000 (p<.001). In all of these bands, the B2 group outperformed the B1 group. Surprisingly, the test was only able to significantly distinguish between B1 and C level readers at bands 3,000 (p<.001), 4,000 (p=.001), 6,000 (p<.001), 8,000 (p=.002), and 10,000 (p<.001). This may, however, be an artefact of the rather conservative Bonferroni correction. For instance, the p-value for the 5,000 band was .013 for the comparison between these two groups. All significant differences in means showed, as expected, a better performance of the more proficient reader group. When comparing scores from B2 readers with C level readers, the test did not manage to distinguish significantly between these two learner groups at any frequency band. Table 40 below summarizes the results of the statistical significance tests for the comparison of group means.

Table 40: Summary of post hoc test results

Band	**A2-B1**	**A2-B2**	**A2-C**	**B1-B2**	**B1-C**	**B2-C**
500			X			
1,000						
1,500		X				
2,000				X		
2,500		X	X	X		
3,000		X	X	X	X	
4,000			X	X	X	
5,000				X		
6,000			X		X	
8,000		X	X	X	X	
10,000		X	X	X	X	

The table illustrates that the vocabulary test does not seem to be able to distinguish clearly between the vocabulary breadth of A2 and B1 learners, and between that of B2 and C level readers. The test manages to distinguish significantly different scores between A2 and B2 learners and between B1 and C level learners at five frequency bands and between A2 and C level learners and B1 and B2 readers at seven out of eleven frequency bands.

7.4. Discussion

The aim of the study presented in this chapter was to explore the relationship between vocabulary knowledge profiles and the reading proficiency levels of EFL learners in order to add to the validation argument of the vocabulary test. Three main findings from the investigation appear particularly relevant.

First, the emerging vocabulary profiles of readers at different levels do mostly correspond to expectations relative to each other. The percentage mean scores per level were generally highest in all levels for the C level readers, followed by B2 and B1 level readers, with A2 and A1 level readers performing worst on the test in the different frequency bands. While readers between A2 and C level all appear to perform almost indistinguishably well at the three highest frequency bands, with A2 readers even outperforming C level readers in band 2, an effect of frequency and proficiency seems to appear as of the fourth band. The fact that the frequency profiles are following a mostly expected pattern seems reassuring for the validity of the vocabulary test. What might be interpreted as worrisome is that the A2 and B1 vocabulary knowledge profiles are very similar, with A2 readers even outperforming B1 readers in three frequency bands. This is the second striking finding from the present study.

The study has shown that the test, in its current iteration, is not able to distinguish with reasonable precision between the vocabulary breadth of A2 and B1 learners. While not ideal, it is less concerning that it also failed to discriminate in any level between the vocabulary knowledge of B2 and C level readers as C level learners are not the target audience for the vocabulary test. It is, however, disconcerting that no significant differences in mean scores were found in the comparison of A2 and B1 readers at any frequency band. The threshold, quite literally in terms of the CEFR, between these two levels is critical. B1 denotes the transition to being an independent user of English (Council of Europe, 2001) and is perhaps one of the key levels to reach for language learners. For most learners, this might be the highest level they may achieve, and for many learners A2 or B1 may be the highest level they need to achieve, e.g. for the purposes of residence or citizenship (e.g. see Council of Europe, 2014; Rocca, 2017). While this may arguably be a weakness of the current test, it needs to be confirmed with more readers of these two proficiency levels. Should a lack of expected differentiation between the typical vocabulary knowledge profiles of these two proficiency groups prevail even with a larger sample, then this is problematic for the test's validity.

What is reassuring about the quality of the vocabulary test, however, is that it generally is successful in distinguishing between B1 and B2 learners. The fact that a significant distinction in group means were found for most of the frequency bands speaks for the validity of the test. It is also noteworthy that this consistent distinction was found between the two largest participant groups, which reaffirms the need for a larger participant sample in the A2 group to hopefully find a similarly strong discriminatory power of the vocabulary test and add to the validity of the test.

It can further be noted that the test manages to distinguish reasonably well between mean scores of readers at non-adjacent proficiency levels. Significant differences in mean scores were found at five to seven bands between A2 and B2 learners, between B1 and C level learners, and between A2 and C level readers. However, these differences have not been as pronounced as expected. Again, though, this could be related to the relatively small proficiency subgroups. Hypothetically, the results could also mean that the vocabulary size differences between CEFR levels are smaller than expected, particularly for high-frequency words (e.g. Hulstijn, Schoonen, De Jong, Steinel, & Florijn, 2012) and that the assumption of a distinct and linear vocabulary size difference between each of the CEFR levels is flawed in principle. Further validation will have to be carried out to confirm or rule out this hypothesis.

It goes without saying that these are only tentative results, and any inferences are severely limited by the small sample size. Additional data will need to be gathered to corroborate the findings. With more test takers, and potentially also a less conservative significance level adjustment for the multiple comparisons, it is speculated that not only will the emerged differences be confirmed, but also that there will be a more precise differentiation between the three lowest proficiency groups of readers (A1, A2 and B1). Adding more lower-level test takers will also clarify whether there are discernible differences already in the first three frequency bands, or whether these more narrow frequency bands hold less diagnostically useful value than anticipated.

7.5.Summary

The study presented in this chapter has analysed the relationship of the test scores at different frequency bands yielded by the newly developed vocabulary test and a standardized measure of reading comprehension. The aim of the research was to investigate the assumption that a highly valid test of written receptive vocabulary knowledge would be able to distinguish clearly in the generated profiles between readers at different proficiency levels. The frequency profiles broadly followed an expected pattern. With very few exceptions the profiles and mean % scores per frequency band were in the expected order of proficiency with higher means for higher proficiency groups. Given the limited size of proficiency subsamples, preliminary evidence was found that the test manages to differentiate between proficiency levels as of the fourth frequency band. However, the test did not consistently perform well in significantly distinguishing between readers of adjacent proficiency levels across all bands. In particular, no difference was found between scores of A2 readers and those of B1 readers. However, this may be due to the small subsample sizes as well as a very conservative significance level adjustment. Additional data is needed before any robust claims about the validity of the vocabulary test can be made and the typical vocabulary knowledge profiles of different reader proficiency groups can be incorporated into the score report and feedback of the test as intended. Nevertheless, the findings do contribute to initial validation evidence that can be integrated into a validity argument of the newly designed test (see Appendices K and L for the full test). The development of an initial version of this argument is the focus of the next chapter.

8. Validity argument

As pointed out in Chapter 2, the validity of a test score interpretation is dependent on the purpose for which a test was designed. As such, it should not be the case that a test is developed and only then explored for the purposes it could be used for, but rather that the purpose of a test needs to be determined at the outset as this has crucial repercussions on the development of the test itself. Taking validity and validation into account from the outset has been termed validity by design (Mislevy, 2007) or design validity (Briggs, 2004).

The present vocabulary test was designed from the start as a diagnostic measure of form-meaning link knowledge of beginner to intermediate international learners of English as a foreign language. The purpose of the test is to generate diagnostic profiles of vocabulary knowledge linked to frequency bands of a large corpus of general English. Harding, Alderson and Brunfaut (2015), for instance, have argued that "a measure of one's vocabulary size and depth would be very useful, especially if [...] results were reported in bands of the frequency levels of the occurrence of words" (p.7). The present test development project is an attempt to begin addressing this gap and provide an additional tool for the assessment repertoire of classroom teachers as well as SLA researchers, particularly one that is useful for diagnosis. Alderson, Brunfaut and Harding (2014) claim that "[t]he ongoing development of testing instruments which target specific, atomistic aspects of language knowledge and/or performance is vital for developing a professionalized system of diagnosis" (p. 22). This chapter will attempt to construct an initial draft of a validation argument to show how this instrument could contribute to this.

Although the field of diagnostic testing is comparatively under-researched and undertheorized (Alderson et al., 2014), Alderson and Huhta (2011) and Alderson et al. (2015) state some tentative characteristics that diagnostic tests should feature to be considered valid. They maintain that diagnostic assessment tools:

- are more likely to be discrete-point and focused on specific elements

- are inevitably less concerned with authenticity
- are typically low- or no-stakes tests
- involve little negative affective barriers that may inhibit performance
- provide results immediately or with as little delay as possible
- take advantage of computer-adaptiveness
- provide detailed analysis and feedback in their score reports that a test taker can interpret meaningfully and act upon
- lead to remediation or further instruction
- are more likely to focus on language than on language skills
- are more likely to focus on low-level language skills than integrated higher-order skills
- are informed by SLA research and theory
- are based on a specific theory of language development or at least on content covered in instruction
- focus on weaknesses rather than strengths

This chapter will outline and recap how these features were considered in the design of the lexical knowledge measurement instrument and will also serve as a checklist in the evaluation and part of the validation of the resulting test. The checklist will be integrated into a validation argument based largely on Bachman and Palmer's (2010) Assessment Use Argument. As already discussed in Chapter 2, this assessment use argument will have to be adapted as it stems from the paradigm of communicative language testing, which is arguably slightly different to diagnostic testing. The framework also has to be viewed critically, as has been outlined in Chapter 2. Particularly in light of Fulcher's (2015) general critique of utilitarian validation approaches, the ultimate goal of beneficial consequences that argument-based validation approaches are often based on needs to be considered with caution. However, the argument structure itself that these approaches employ is still viewed as useful in validation, even by critics of the AUA such as Fulcher (Fulcher, personal communication). Since it is beyond the remit of the present thesis to develop a completely new validation framework that is based on the notions and principles of Pragmatic realism, this chapter attempts to appropriate an existing framework as much as necessary or possible to probe whether this validation framework is a viable way forward for guiding validation efforts of vocabulary tests in general, and diagnostic vocabulary tests in particular. This seems particularly relevant as there have not been any attempts to validate vocabulary tests using an argument-based validation scheme, with the exception of Voss' (2012) computer-based ESL academic collocational ability test. He bases the validation of the collocation test developed in his thesis on Kane's (2001) validity argument, which underlies Bachman and Palmer's AUA.

In the following, therefore, the evidence gathered in the studies described in this thesis will be compiled and integrated into the blueprint of an assessment use argument. Wherever applicable and available, a case will be built for the validity of the test, while also acknowledging where additional research will be necessary before the test can be launched for public use. In doing so, it will pinpoint to the studies and steps that will be undertaken in the future because they were outwith the remit of this PhD. This will pertain to the section of consequences first and foremost. Warrants for which empirical backing is still (at least partially) outstanding will be marked with *.

First, however, the structure and elements of an assessment use argument need to be detailed. Bachman and Palmer (2010) specify that an AUA is a series of inferences to link a "test taker's performance to a claim about assessment records, to a claim about interpretations, to a claim about decisions, and to a claim about intended consequences, along with warrants and backing to support these claims" (p. 103). The argument therefore essentially consists of five elements: Consequences, decisions, interpretations, assessment records and the candidate performance. Four of these stages Bachman and Palmer label as claims, which in turn are supposed to satisfy certain quality criteria. Consequences of an assessment use are supposed to be beneficial to stakeholders. Decisions need to be value sensitive and equitable. Interpretations have to be meaningful, impartial, generalizable, relevant and sufficient for the decisions that will be made on the basis of the assessment. Assessment records should be consistent. The general structure of an AUA is illustrated in the figure below:

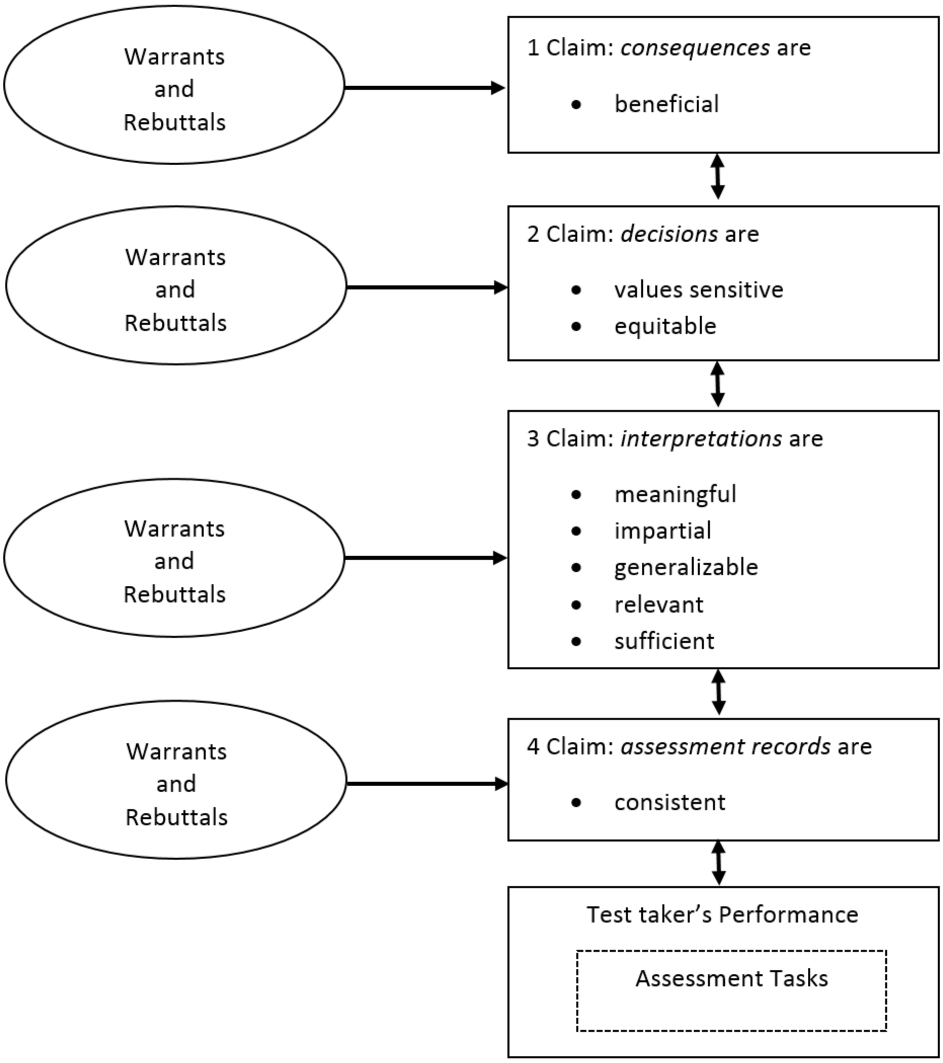


Figure 31: Claims and warrants in an Assessment Use Argument (Bachman & Palmer, 2010, p. 104)

Any of these claims can be adapted to a specific assessment situation and have warrants associated with them, describing the quality of a particular claim. Theoretical and empirical backing needs to be put forward in the argument to support or rebut these warrants. While Bachman and Palmer (2010) provide an illustrative list of warrants for the four claims, they also acknowledge that their list will neither be comprehensive nor will all warrants require addressing in all assessment situations. In the following, therefore relevant warrants will be selected and adapted for the present assessment use argument.

8.1.Intended consequences

Claim I: The consequences of using the Vocabulary Knowledge Profiler and of the diagnostic decisions that are made based on its results are beneficial to the EFL learners and EFL teachers using the test in their classroom learning and teaching.

**Warrant A1: The consequences of using the Vocabulary Knowledge Profiler that are specific to the test takers and teachers will be beneficial.*

Warrant A2: Assessment reports from the Vocabulary Knowledge Profiler are treated confidentially.

**Warrant A3: Assessment reports from the Vocabulary Knowledge Profiler are presented in ways that are clear and understandable to the test takers and the teachers.*

Warrant A4: Assessment reports from the Vocabulary Knowledge Profiler are accessible to stakeholders immediately upon completion of the test.

**Warrant A5: Use of the Vocabulary Knowledge Profiler will help to promote good instructional practice in language teaching and effective language learning by linking the vocabulary profiles to both coverage research findings and CEFR levels for reading.*

**Warrant B1: The consequences of the EFL learner's changes in their vocabulary and reading learning practice based on their self-diagnostic decisions will be beneficial to the students.*

**Warrant B2: The consequences of the EFL teacher's changes in their instructional practice based on the teacher's diagnostic decisions will be beneficial to the students.*

**Warrant B3: The consequences of the EFL teacher's changes in their instructional practice based on the teacher's diagnostic decisions will be beneficial to the teacher.*

Even though it will hopefully also be used by SLA researchers, the main stakeholders involved in the use of the Vocabulary Knowledge Profiler will be EFL learners from all across the globe and their EFL teachers. Following the principles listed by Alderson & Huhta (2011), the test is not intended as a high-stakes test, or even as a proxy for a proficiency indication. Hence, it is unlikely that there will be many other stakeholder groups affected. Even though parents of EFL learners, for instance, could hypothetically be indirectly affected by the assessment use, their involvement appears negligible in the low-stakes scenarios for which this profiler was designed. The profiler will be made freely available to them online and will hopefully help inform their instructional decisions. The intended consequences are therefore that, in providing an up-to-date and easily accessible tool that is in some principles similar to the Vocabulary Levels Test, the Vocabulary Knowledge Profiler will be useful to learners and teachers in performing some of the same purposes as the VLT. The intention is for the profiler to guide learners in their learning of vocabulary by identifying weaknesses in lexical knowledge in different frequency bands. It is hoped to be beneficial for their vocabulary learning by allowing them to graphically recognize which frequency bands they have already mastered relatively well, and which frequency bands require attention in future learning. As a self-assessment tool, it should further be useful for learners to interpret their vocabulary knowledge profile vis-à-vis typical profiles of other learner groups at different proficiency levels. The study presented in Chapter 7 of this thesis is a first step towards this, but further evidence will have to be collated before this feature can be implemented in the score report with confidence. The fact that after careful consideration and empirical investigation an algorithm was chosen that terminates the test before any test takers is exposed to too many difficult words should further benefit the learning of candidates as it hoped that this approach will have motivational benefits.

For teachers, the use of the Vocabulary Knowledge Profiler should allow for better diagnostic decisions about the lexical abilities of and future instruction needs for learners. The group report feature is also hoped to be useful for group level instructional decisions, such as determining the suitability of and selecting reading materials for students. This feature is already implemented and allows teachers to create a user account and generate as many ID codes for their students as they wish. These ID codes can then be used by their students and the system automatically recognises them as belonging to this teacher's group and sends a score report of that ID to the teacher's account.

In theory, therefore, there is reason to believe that this first claim and warrant A1 are supported. Transparent, detailed, immediate and encouraging feedback provided to teachers and learners through a low-stakes profiler, related to descriptors of what learners at different levels *can* do should enable positive self-evaluation and learning appraisal (Dörnyei, 2001), add the motivational benefit of reducing anxiety (Dewaele & MacIntyre, 2014; Dörnyei & Ryan, 2015; Simsek & Dörnyei, 2017), and hopefully even facilitate a successful pathway of a "directed motivational current" (Dörnyei, Ibrahim, & Muir, 2015; Dörnyei, Muir, & Ibrahim, 2014) for the learners. Although the stakeholders were not directly involved in the test development, feedback from stakeholders throughout trial administrations has been encouraged and integrated in the current iteration of the test, such as the option to view which items have been answered correctly and which have been answered incorrectly, or the feature of a coloured frame appearing around the selected answer in the test so that candidates can see whether the computer has accepted their click. Empirical backing for this first warrant, however, is still outstanding. As Bachman and Palmer (2010) note, for low-level assessments such as this Vocabulary Knowledge Profiler, simply articulating warrant A1 may be convincing enough "to primary stakeholders without the need for extensive backing" (p. 181), so the professional knowledge and test development experience of the test developers may suffice as evidence to some test users. However, in the case of this instrument, a survey of users is planned to evaluate the perceptions of the profiler as well as its score report once the abovementioned feature has been implemented. Self-report data about the usefulness of the diagnostic report needs to be gathered to provide empirical backing for this warrant of beneficial consequences. In addition to this survey, a study will have to be devised, in which the actual language learning is monitored alongside the self-report data as relying solely on self-report data would be insufficient to build a strong assessment use argument.

The backing for warrant A2 can be provided without additional research. All test taker responses are recorded on a private password-protected server and with no personal data that would allow identifying any particular candidate. At no point is the identity of a test taker known to the operator of the profiler. ID codes are generated to protect the anonymity of learners. Only with teacher-generated and teacher-distributed ID codes is the person behind the ID code known to that one teacher. The profiler itself only collects minimal biodata on a voluntary basis: age, gender, years of learning English, and L1. This way it is ensured that this warrant is implemented and assessment records are "provided only to the test takers themselves and individuals who are authorized to receive them" (Bachman & Palmer, 2010, p. 183), i.e. their teachers. For research use of the data, informed consent will be obtained by the test developer before any anonymous use.

As for warrant A3, the test designer has taken and is planning to take a number of measures to ensure that assessment reports from the Vocabulary Knowledge Profiler are presented in a clear and understandable manner to test users. The graphical score report should be very accessible, particularly once the option of superimposing other proficiency groups' profiles over one's personal profile is implemented. This link of vocabulary test scores to language use, in this case reading, and the Common European Framework level descriptors in particular should, in theory, add value to the score interpretation. It will further be necessary to make user's manuals available, for at least three different audiences: learners, teachers, and researchers. Particularly learners and teachers will hopefully find instructions in plain and non-technical language helpful on how, when and why to use the test and how to interpret its scores in relation to CEFR proficiency levels and coverage research. The use and usefulness of these user manuals, however, will in turn have to be monitored and researched so as to make sure they are indeed clearly understandable and useful for diagnostic feedback. Despite the widespread use of the VLT, users may not be familiar with either up-to-date coverage research (Schmitt et al., under review) or the proficiency framework of the CEFR and its rich descriptors, so these manuals for different audiences will attempt to aid in making the feedback "as relevant, complete, and meaningful to the test taker as possible" (Bachman & Palmer, 2010, p. 184). In addition, information for researchers will be provided – ideally in the form of a peer-reviewed publication - on the test website with technical details of the profiler's construction and validation to allow for scrutiny, potential replication and feedback from the scholarly community so as to continually improve the instrument.

Warrant A4's backing is the evidence that assessment reports from the Vocabulary Knowledge Profiler are indeed presented to test takers immediately upon completion of the test. Also, for ID codes generated by teachers, these score profiles are sent to the teacher's account automatically once the test is completed.

As with warrant A1, empirical evidence to back warrant A5 still needs to be generated. Given the history of use of the VLT and the intention that this new profiler will provide a similar but updated instrument, it is assumed that use of the Vocabulary Knowledge Profiler will help to promote good instructional practice in language teaching and effective language learning by linking the vocabulary profiles to both coverage research findings and CEFR levels for reading. An impact or washback study of the profiler's use is therefore necessary. A potential rebuttal to the warrant is that the test, in its current iteration, does not include any type of formulaic sequences and focuses on single word units instead, which may lead to negative washback in terms of students learning only individual words. This, alongside the fact that this test presents items in a discrete and largely decontextualized manner, could be a point of criticism when considering that words normally occur in context, which usually also impacts their precise meaning and needs to be taken into account in vocabulary learning (Schmitt, 2010). However, as outlined in Alderson and Huhta's (2011) principles, diagnostic tests are by definition less authentic and focus on rather narrow, or even atomistic, constructs or specific language elements and hence often need to be discrete-point and limited in context. It is further anticipated that the low-stakes nature of this assessment will have an impact on which words are being learned, i.e. which frequency bands, but not necessarily on how these will be learned, i.e. in a completely decontextualized manner. The modular nature of the bespoke computer-adaptive system in the backend of the test (i.e. the operational software and programming) also allows for integrating multi-word units in the future, which should further add to beneficial washback in vocabulary learning.

In the course of such a washback study, evidence will also have to be gathered to support warrants B1, B2, and B3 to monitor positive effects on learning and instruction. The usage statistics of online tools such as text profilers (e.g. www.lextutor.ca) and the international prominence of the VLT shows that teachers are in need of tools to guide their reading and vocabulary instruction. Harding, Alderson and Brunfaut (2015) also maintain that tools such as the Vocabulary Knowledge Profiler can be useful for learners and teachers and it is ultimately these two stakeholder groups that will act as diagnosticians (Alderson et al., 2014) on the basis of the assessment information. Adjustments to instruction are therefore limited to these two key groups, which is reflected in these warrants. There is no rebuttal needed for classification errors and their potential detrimental consequences as the profiler does not attempt to classify learners but instead is intended to help guide decisions to modify teaching and learning based on the score feedback. Although the profiler will attempt to mitigate general measurement error in the scores, the low-stakes nature of the test does not make this a concern for the consequences of the test use.

8.2.Decisions

Consequences, intended or not, result from actions taken by stakeholders on the basis of decisions that are made by test users. The decisions made on the basis of interpretations of Vocabulary Knowledge Profiler scores will unequivocally be low-stakes without exception. In general, these decisions will affect learners and teachers, and potentially researchers should they choose to include vocabulary knowledge profiles in their investigations. Low-stakes decisions, by definition, have relatively minor consequences both at macro and micro level. As such, the present test satisfies yet another criterion of a useful diagnostic test as stipulated by Alderson & Huhta (2011). Following this, decisions based on Vocabulary Knowledge Profiler score reports will be easily reversible. Misdiagnoses and errors in instruction adjustment through assignment of inappropriate learning activities can be corrected relatively quickly once noticed. Since the Vocabulary Knowledge Profiler also allows for self-assessment, the responsibility for the diagnostic decision may be shared between learners and teachers. Information from the Vocabulary Knowledge Profiler is supposed to provide a basis for decision-making about learning and instruction, by focusing on specific lexical areas of strengths and weaknesses (Alderson et al., 2015). Decisions made by teachers may lead to changes in learning activities, materials provided or even syllabi. Bachman and Palmer (2010) note that "the teacher may provide diagnostic feedback to students based on their performance on the assessment, and suggest that they focus on specific areas on language ability in which they need to improve" (p. 197). This is exactly the kind of formative decision that should be facilitated through the interpretation of Vocabulary Knowledge Profiler scores. When used as a self-assessment by students, students may themselves decide autonomously on future learning activities and goals. While the Vocabulary Knowledge Profiler may also be useful for placement purposes, the current validation effort and, more importantly, the present assessment use argument does not speak to this purpose and would require additional validation studies. The Vocabulary Knowledge Profiler was also not primarily designed to enable decisions about learner progress as it will likely be too coarse a measure to pick up on incremental changes in vocabulary knowledge over short instructional periods. This is where the assessment use argument for this particular instrument needs to be adapted quite considerably as there are no categorization, classification or certification decisions involved in the use of this diagnostic tool. Hence, concerns for equitability and value sensitivity are less serious than for achievement or other higher-stakes tests. Warrants pertaining to "the relative seriousness of false positive and false negative classification errors" (Bachman & Palmer, 2010, p. 201), decisions and communication about cut scores and remedying classification errors, equal opportunities to acquire the ability assessed for achievement or certification, or the ruling out of any other considerations in classification decisions except the cut scores and decision rules simply do not apply in this case. While all of these issues and warrants require addressing with backing and rebuttals in scenarios of placement, course admissions, or general high-stakes proficiency testing with usually many stakeholder groups affected, they are mostly irrelevant for instruments intended for formative or diagnostic use. For the intended purpose of arriving at useful diagnostic decisions, however, the following argument is proposed.

Claim II: The diagnostic decisions that are made on the basis of the interpretation take into consideration existing educational and societal values and relevant laws, rules, and regulations, and are equitable for EFL learners.

Warrant A: Relevant educational values of teachers, and (if applicable) school regulations, are carefully considered in the decisions to modify instruction based on the diagnostic information.

Backing to support warrant A is not easy to provide as it seems highly context-specific and the decisions based on the use of the Vocabulary Knowledge Profiler are made not by the test developer but by teachers and students themselves. It is therefore assumed that using the instrument implies certain educational values that are shared with the test developer, such as the importance of formative feedback based on solid diagnosis to maximize the effectiveness of foreign language pedagogy, but ultimately this seems beyond the realm of what the test developer can influence in this case. Similarly, the implicit expectation of parents or legal guardians of learners in language learning institutions that teachers will try as best they can to monitor, evaluate and improve the effectiveness of their instruction is considered but virtually impossible to integrate into a validity argument of this kind. The user manuals will certainly highlight the added value provided by the use of the tool and will try to guide teachers and students towards appropriate decisions based on their outcomes, with suggested courses of action exemplified as to how instruction can be modified in line with the test designer's intentions. Given that the profiler is intended for use all over the world in different educational contexts, the abidance by local school regulations or other legal parameters is also outwith the reach of the test designer. It has been designed with due diligence not to violate any ethical or legal requirements, and it appears difficult to imagine a scenario where use of this tool for its intended purpose would clash with any laws or educational values. As with most warrants in this stage of the assessment use argument, it needs to be stressed that the low-stake nature of the classroom- or self-assessment use of this tool is quite distinct from assessments that determine whether or not particular candidates will have access to certain resources, educational or otherwise. While it is certainly the case that in such scenarios educational and societal values may not always be consistent with prevalent laws and regulations or that values of different stakeholder groups might even be competing (Bachman & Palmer, 2010), this is generally not anticipated to be an issue with diagnostic assessments.

8.3.Interpretations

While intended beneficial consequences and relevant decisions need to be thought about and articulated at the start of the test development process, the interpretations of test scores are really what is at the heart of the assessment use argument and thus also test validation. They provide the information needed to make appropriate decisions and thus also constitute the major part of argumentation in the case of the Vocabulary Knowledge Profiler with most research to back warrants up to now having gone into supporting this claim of the argument.

Claim III: The interpretations about the written receptive vocabulary knowledge at the form-meaning link level assessed in the Vocabulary Knowledge Profiler are:

- *meaningful with regards to the vocabulary knowledge needed to be employed when performing reading tasks and with respect to general SLA theory of vocabulary learning,*

- *impartial to all groups of test takers,*
- *generalizable to subsequent learning activities,*
- *relevant to the formative decisions to be made, and*
- *sufficient for the diagnostic decisions to be made.*

Warrant A1: The construct of the Vocabulary Knowledge Profiler is written receptive form-meaning link knowledge of the most frequent 10,000 content lemmas of a representative contemporary corpus of general English. This construct definition is based on coverage research from SLA vocabulary acquisition theory.

Warrant A2: The test specifications clearly specify the administration and computer-adaptive algorithm as well as the scoring of the Vocabulary Knowledge Profiler. The conditions under which learners complete the test and how their answers will be elicited is clearly laid out so that inferences about the assessed construct can be made.

Warrant A3: The procedures for administering the assessment enable test takers to perform at their highest level on the ability to be assessed.

Warrant A4: The procedures for producing the assessment record focus on the aspects of the performance relevant to the assessed construct.

Warrant A5: The assessment task, i.e. the item format, engages the written receptive form-meaning link vocabulary knowledge of candidates as specified in the construct.

Warrant A6: Assessment records, i.e. the vocabulary knowledge profiles generated, can be interpreted as indicators of written receptive form-meaning link vocabulary knowledge.

**Warrant A7: The test developer will communicate the construct definition in terms that are clearly understandable to all stakeholders.*

**Warrant B1: The item format or item content do not favor or disfavor a particular subgroup of test takers.*

Warrant B2: The test items do not include content that may be topically or culturally offensive or linguistically inappropriate to some test takers.

**Warrant B3: The procedures for producing an assessment record are clearly described in terms understandable to all test takers.*

Warrant B4: Individuals are treated impartially during all aspects of test administration, including equal access in terms of cost, location, familiarity with equipment, as well as equal access to information about assessment content and procedures. They also have equal opportunity to demonstrate their written receptive form-meaning link vocabulary knowledge.

**Warrant B5: Interpretations of the written receptive form-meaning link vocabulary knowledge of candidates are equally meaningful across all groups of test takers.*

Warrant C: The characteristics of the setting, input and expected response do not correspond to tasks usually found in the target language use domain due to the diagnostic nature of the test. However, the scores in each band do allow to generalize to a given frequency band, the contents of which are sampled from authentic materials of the target language use domain. There is also an established link between the vocabulary knowledge profile and CEFR reading levels from a proficiency test whose task characteristics resemble more closely those of target language use tasks.

**Warrant D: The assessment-based interpretations provide information relevant to the diagnostic decision-making. The information yielded is helpful for learners and teachers in planning future instruction and learning activities.*

**Warrant E: The assessment-based interpretations provide sufficient information for the diagnostic decision-making. The Vocabulary Knowledge*

Profiler offers enough information about test taker's mastery of each frequency band to make an informed diagnostic decision.

The research in this thesis has mainly aimed at collecting backing evidence for the consistency and meaningfulness of score interpretations for this new instrument. Based on Nation's (2001) taxonomy, it has delineated its construct clearly as written form-meaning link knowledge of vocabulary knowledge enabling reading. While not a clearly psychometrically distinct and separable dimension of vocabulary knowledge (González-Fernández & Schmitt, under review), this aspect of vocabulary knowledge has been pointed out by various theorists as the key element in vocabulary learning and the one most crucial for reading comprehension (Laufer & Goldstein, 2004; Schmitt, 2010). This construct is operationalized in the recognition format of a four-option multiple-choice item type, in which a target L2 word form of a single word is presented and has to be matched to a picture, L2 synonym or L2 definition corresponding to the most frequent meaning of that word form. While using a meaning recognition format, the construct of written receptive vocabulary knowledge for reading implies a link to knowledge at the meaning recall level. In reading, a learner is confronted with the L2 form and has to recall, without any help other from the context, i.e. no alternative meaning options, the meaning of that particular word form. The construct has been shown to be unidimensional in the IRT analyses of the item piloting (see Chapter 5). Following Schmitt's (2014) argument, it is mainly a profiler of vocabulary breadth to the minimal depth level of the form-meaning link.

The construct is further specified in the test specifications as comprising knowledge of the most frequent 10,000 content lemmas of a representative contemporary corpus of general English. The decision to operationalize a frequency-based vocabulary test and employ frequency-based item sampling and score reporting is informed by existing SLA research that has confirmed frequency as a key driver of acquisition, not just of lexis (Ellis, 2002; Nation & Webb, 2011). It has also been motivated by enabling a connection to the established field of coverage research, which has indicated that mastery of particular frequency levels enables learners to perform specific linguistic tasks such as reading authentic novels or viewing television programs (Schmitt et al., under review). The diagnostic meaningfulness of the construct definition is based on this coverage research that has been carried out for and with EFL learners. The figures proposed for reading activities (Schmitt et al., under review) provide the frame of reference for the construct to guide diagnostic decisions.

The counting unit of the specified construct has been determined as lemmas, specifically content lemmas, as they have been shown to be more psycholinguistically valid for EFL learners and provide more meaningful interpretability than previously used level 6 word families (Bauer & Nation, 1993). Item sampling for this profiler is therefore based on lemmatized frequency lists from the Corpus of Contemporary American English (COCA) (Davies, 2008-), which was selected as a state-of-the-art large-scale balanced corpus of general English representative of what learners of English might or wish to encounter in authentic discourse. The COCA provides part-of-speech-tagged frequency lists which include distribution information and is probably the standard reference corpus available. The reporting in frequency bands is also based on the word list from this corpus, thus providing the user with scores that can be meaningfully interpreted on the basis of a corpus of contemporary English, which makes the Vocabulary Knowledge Profiler's construct definition and item sampling one of the most modern of the vocabulary tests available. In addition, the conceptualization of the Vocabulary Knowledge Profiler as a computer-adaptive test with a bespoke item database system in the backend of the test would allow in the future to update frequency information or change frequency-band categorization tags with minimal effort. This means that the system can easily accommodate for changes in rolling corpora, or even integrate frequency information from more suitable or up-to-date databases should they arise. As such, the sampling and reporting of the Vocabulary Knowledge Profiler can constantly be ensured to correspond to the vocabulary found by learners in the target language use situation.

The study presented in Chapter 7 of this thesis has further attempted to enhance the meaningfulness of score interpretation and generate backing for warrant A1 by trying to establish a link between score profiles and CEFR proficiency levels in reading. While further validation research with a larger candidature will be necessary to confirm this link, particularly at the lower proficiency levels, the initial results are reasonably promising for the B1 and B2 levels given the strict significance adjustment. Ultimately, learners will be able to call up average profiles of readers at different proficiency levels to compare their scores across the frequency bands with learners from those groups, whose reading ability is described in detail in the CEFR descriptors. This will allow users to identify lexical gaps and deviations from those average profiles so that they can easily diagnose weaknesses and recognize which frequency bands will require attention in their learning going forward.

Another measure that has been taken to facilitate score interpretation and make the feedback more useful and meaningful is the new approach to frequency banding implemented in the Vocabulary Knowledge Profiler. As described in detail in Chapter 4, these finer-grained frequency bands at the high frequency end and the broader bands at the lower-frequency end are both backed by theoretical considerations as well as empirical data from corpus analyses. These have shown that the power of frequency as a clustering factor decreases along the continuum and that the relative importance of frequency bands in terms of coverage should be considered in sampling items for diagnostic vocabulary tests.

The backing for warrant A2 is that the test specifications presented in Chapter 5 clearly specify the administration and computer-adaptive algorithm as well as the scoring of the Vocabulary Knowledge Profiler. Items are presented as four-option multiple choice items with an "I don't know" option to skip the item. The item stem provides the target word and a short non-defining context sentence in the target language to indicate the word's part of speech. Options are presented as stock images or synonyms and short simple definitions in high-frequency language depending on the frequency band. All options are words from the same part of speech and the same frequency band and are plausible in the sentence context. One of the four options is correct. The selected computer-adaptive algorithm has been shown in Chapter 6 to produce representative and reliable results. Even though the item selection for each test version is randomized and the number of items presented varies depending on candidates' answers, the administration itself is standardized through the computer delivery. The test conditions are clearly stated so there should be little to no interference with the measured construct. This being said, it is impossible to rule out some test-taking strategies, test-wiseness or even guessing completely given the item format employed. These factors would confound the scores as they introduce construct-irrelevant variance and measurement error. In Chapter 4, it was argued that a correction or adjustment formula should be considered given the findings of the study presented. The study, however, only examined test takers from one L1 background at a relatively homogeneous and proficient level. It will therefore be necessary to conduct another investigation with a more diverse learner group to find an appropriate adjustment formula that will then be implemented to account for these factors to some extent. Part of such a study will also be to probe at which level of inaccuracy such an adjustment formula should be factored in. The computer delivery system also allows for flagging conspicuous response patterns. In this way, candidates that are, for instance, always selecting the option displayed in the top right corner can be identified and their answers invalidated even though they might score enough points by chance to proceed through a number of frequency bands.

The backing for warrant A3 is that the paradigm of "biasing for best" (Swain, 1983) has been taken into account throughout the entire test development process. The Vocabulary Knowledge Profiler with its "floor first" algorithm elaborated in Chapter 6 allows learners to proceed through mastered frequency bands quickly and terminate the test before encountering too many unfamiliar, difficult and potentially demotivating items. At the same time, it probes with additional items those frequency bands that are of particular interest to the learners because of identified weaknesses and provides a score report that gives immediate and useful feedback by allowing the candidate profile to be compared against profiles linked to can-do descriptors of CEFR reading levels. An example item demonstrating the item format is provided in the form of a short video clip. Candidates can take as long as they need to complete the test as it is not taken under any time constraints. However, a recording of the timing per item would be possible through the computer-adaptive system in the future if speededness and fluency/automaticity of access was desired to be incorporated into the construct for a particular purpose. If students are familiar with computerized test taking in general, which more and more students are, then this testing environment should feel familiar and comfortable to candidates.

The Vocabulary Knowledge Profiler is scored automatically by the computer system according to a predefined scoring key. The dichotomous scoring procedure and the unambiguity of the correct items has been reviewed by native speakers and experts in the field. The study presented in Chapter 3 of this thesis has provided evidence that the item format represents the construct of word knowledge reasonably well and, with some limitations, allows for inferences about the true word knowledge of candidates. This evidence also serves as backing for warrant A5. Although not directly engaging learners in the task of meaning recall because of its impracticalities for testing, the item format has been established as tapping into the written receptive form-meaning link vocabulary knowledge of candidates. The aforementioned follow-up study will, however, need to monitor closely test taking strategies and guessing through think-aloud protocols and post-test interviews or written meaning recall measures so that these findings can be corroborated and incorporated into adjustment formula if necessary. As Bachman and Palmer (2010) rightly note: "just believing that an assessment task engages the ability to be assessed is not enough evidence to support this warrant" (p. 228). Given the present research outlined in Chapters 3 and 6 of this thesis, however, there is some evidence that the vocabulary knowledge profiles can be interpreted as indicators of written receptive form-meaning link vocabulary knowledge at the different frequency levels. Particularly further data on the relationship between vocabulary profiles and CEFR reading levels will be useful "evidence of convergence" (Bachman & Palmer, 2010, p. 229). Even more so, if several different proficiency tests' scores can be used for these studies. Correlation studies with existing tests of vocabulary breadth could also be conducted, although it must be noted that the slight difference in construct, counting unit and frequency banding might render comparisons problematic for validation purposes.

Backing for warrant A7 is not available yet, but will be addressed through the production and piloting of user manuals for learners and teachers, as outlined above. This will make sure that test purpose, test construct, test procedure and score interpretation are communicated in a clearly understandable way to all stakeholders. Feedback on the user-friendliness of the test and the manuals will be obtained in the course of this so as to maximize the diagnostic usefulness of test use and score reports for all users.

Regarding warrants B1 and B2, it can be stated that target items were sampled randomly from each frequency band from a general English corpus and therefore item content may not favor or disfavor particular individuals or groups of test takers. The fact that a corpus of American English was used may cause minimal additional challenges for students that have been taught British English spelling. However, since the word form does not need to be produced in the selected item format, this effect is likely to be negligible and the global influence of American English in ELT appears to justify the reference corpus selection. Great care has also been taken to avoid topically or culturally offensive items or distracters, particularly in picture-based items, and linguistically inappropriate items. This has been reviewed and checked repeatedly by vocabulary assessment experts. Moreover, no inappropriate items have been reported to the test developer from the administrations to date, which have – in total – been fairly large-scale and heterogeneous in terms of cultural backgrounds (see Chapter 5 in particular, as well as Chapters 6 and 7).

The response format is one of the most widespread item formats and can therefore be assumed to be familiar to the majority of test takers. For those test takers that are not familiar, the example item at the beginning of the test is included. The format itself, however, will require further attention in light of warrant B1. A differential item functioning (DIF) analysis will have to be carried out to ensure that there is no gender (Takala & Kaftandjieva, 2000) or other bias prevalent. Particular attention will also be given to cultural group differences (albeit based only on L1 categorizations). DIF analyses establish whether examinees of the same ability level but from two different groups have different probabilities of answering an item correctly. A test free of items that show DIF will constitute strong backing for warrant B1.

Warrant B3's backing is again covered in the user manual that will be provided to the different stakeholder groups. In this document, the scoring procedures, among other things, will be explained in accessible terms. Test takers and teachers will be asked to review these manuals to make sure that the wording is understandable. For researchers, there will be a more technical report on the scoring procedure available. This provision of information then also functions as backing for warrant B4, which states that test takers are treated impartially during all aspects of test administration, including equal access in terms of cost, location, familiarity with equipment, as well as equal access to information about assessment content and procedures. The Vocabulary Knowledge Profiler will be freely available online to all test takers, as will be the information about the instrument. The administration and scoring is completely anonymous and done by machine so the test will be treating all test takers impartially. Aside from varying stability and speed of internet connections worldwide, administration conditions will be identical for all test takers. However, the random item selection per band and particularly the randomization of the response orders within any one item will need to be monitored so as not to disadvantage individual students.

Backing for warrant B5 comes from the piloting and research studies presented in this thesis (Chapters 4, 5, 6, and 7) and will accumulate as more test takers from more L1 backgrounds use the instrument. The score interpretation is meaningful because a) the item format has been shown to correspond reasonably well with the meaning recall knowledge of the form-meaning link as verified in interviews and written meaning recall measures, b) the counting unit of the lemma allows for better interpretation than the word family, c) the items in the computer-adaptive algorithm are representative of each frequency band, and d) there is a tentative linkage between the vocabulary knowledge profiles and CEFR reading proficiency levels.

As stated in warrant C, the item type does not reflect a typical TLU task. This lack of authenticity is due to the diagnostic purpose of the instrument (Alderson & Huhta, 2011). The study presented in Chapter 6 of this thesis, however, does imply that the scores in each band are generally a robust representation of the overall knowledge of a particular frequency band. Since the contents of these frequency bands are sampled from authentic materials of the target language use domain, i.e. an up-to-date corpus of general English, the scores are generalizable. With the limitation of multi-word sequences and pragmatic nuances of meaning, the lexis encountered in the test corresponds closely to the lexical input encountered in real-life TLU texts. The profiles also allow for generalization to future learning activities as they pinpoint areas of lexical weaknesses. The findings presented in Chapter 7 of this thesis further provide evidence that there is a tentative established link between the vocabulary knowledge profiles and CEFR reading levels. These reading proficiency levels were gleaned from a standardized proficiency test whose task characteristics resemble more closely those of target language use tasks, thus further enhancing the generalizability of the scores and expanding the score interpretation to actual employment of lexical knowledge in skill use. Additional data will be needed to confirm a meaningful link between reading test scores and the vocabulary profiles of the Vocabulary Knowledge Profiler. This should be done by both expanding the study presented in Chapter 7 as well as comparing the ability of the VKP to explain variance in reading comprehension scores with the explanatory power of an existing, concurrent vocabulary measure such as the VLT or the VST.

The decisions based on the interpretations of scores from the Vocabulary Knowledge Profiler can be twofold. One, they can pertain to identifying lexical weaknesses and providing feedback to learners about which frequency bands require attention in future learning to achieve certain thresholds associated with particular tasks or proficiency levels. Two, they can relate to learning activities in terms of selecting appropriate materials that are challenging but within reach for a learner. For either of these two decisions, the information provided in the vocabulary knowledge profiles is relevant as it allows for inferences about mastery at different frequency bands. In this respect, the new instrument is not dissimilar from the Vocabulary Levels Test and its original intention (Schmitt et al., 2001) and the usefulness of this type of information is therefore well-documented. Moreover, experts in the field of both language assessment as well as L2 vocabulary studies have highlighted the relevance of the type of information that this instrument supplies (Alderson et al., 2015; Harding et al., 2015; Schmitt, 2014; Schmitt & Schmitt, 2014). It is therefore anticipated that the instrument will be helpful for learners and teachers in planning future instruction and learning activities. As outlined in the section on Claim I, however, a washback study will have to confirm this empirically.

Sufficiency is a final key criterion for the information providing the basis for score interpretations. The computer-adaptiveness of the Vocabulary Knowledge Profiler implies that test takers will be exposed to a varying number of items per frequency band and in the test in total. In some frequency bands it may be as few as five items that candidates are answering, which could be taken as a rebuttal to warrant E. Indeed, the research literature is inconclusive on this issue with some scholars claiming that five items per band of 1,000 could be enough (Beglar, 2010; Coxhead, Nation, & Sim, 2015), and others arguing that up to 30 items per 1,000 are needed to provide a more accurate estimate (Gyllstad et al., 2015). The research presented in Chapter 6 of this thesis appears to suggest that the number of items selected through the computer-adaptive algorithm of the Vocabulary Knowledge Profiler is sufficient to provide a representative estimate of the frequency band mastery. This may be partly because of the approach to frequency banding. The frequency bands that learners will most likely be doing well and so only encounter 5 items from, are the narrow high-frequency bands of 500 to 3,000. Compared to other existing texts, this corresponds to 10 items per 1,000 if added up to the banding that tests such as the VST use. The Vocabulary Knowledge Profiler acknowledges that this is the minimum number of items, so in reality, most learners are exposed to 10 or even 15 items per frequency band. This, in total, is then comparable to the suggestion by Gyllstad et al. (2015) of sampling 30 items per 1,000. While further research will certainly have to explore whether this is the optimal number of items for each round and thus the test overall, the preliminary evidence appears to provide backing that the Vocabulary Knowledge Profiler offers sufficient information about a test taker's mastery of each frequency band to make an informed diagnostic decision, particularly given the low-stakes nature of the instrument.

8.4.Assessment records

In order to arrive at useful interpretations, Bachman and Palmer (2010) state that high-quality assessment records are a prerequisite. The quality of assessment records is generally a function of their consistency. The following claims and warrants are stated for the Vocabulary Knowledge profiler's assessment records, partly drawing on the warrants explicated in the interpretations section above:

Claim IV: Assessment records of the Vocabulary Knowledge Profiler are consistent across different administrations and across different test taker groups.

Warrant A1: Administration procedures are followed consistently across different administrations and groups of test takers.

Warrant A2: The assessment records are produced automatically by a computer system based on clear specifications.

Warrant A3: Scores of the Vocabulary Knowledge Profiler are internally consistent and scores from different test forms and administrations are equivalent and consistent (reliability).

The test administration and the procedure for producing assessment records is consistent for the Vocabulary Knowledge Profiler as it is a computer-delivered test with no need for human invigilation. Even though the item selection is randomized in each test form and administration, the items are all selected in a prespecified manner from a pool of functioning items. Every test form is therefore generated from the same set of specifications with no interference through human judgment (see Chapter 5 for detailed test specifications). Scoring is completely automatized as specified in an algorithm which scores all tests on identical parameters. The study presented in Chapter 6 of the thesis confirms the test-retest-reliability of the instrument. This type of consistency was identified as the most meaningful for tests of this kind. The IRT analyses from the item piloting demonstrated high reliability of the items, even though internal consistency, particularly in the form of a calculated Cronbach alpha value, was deemed a somewhat problematic concept in vocabulary tests and therefore an expendable index. The consistency of assessment records across different test taker groups will have to be evaluated in the future with a large and diverse candidature in terms of L1 backgrounds. The impact of cognates has been highlighted as a concern in vocabulary tests (e.g. Laufer & McLean, 2016), but might be less of a concern for the diagnostic classroom use anticipated for this instrument. It can therefore be concluded that there is reasonable backing for the consistency of assessment records generated by the Vocabulary Knowledge Profiler.

9. Conclusion

The aim of the present PhD project was to develop a new diagnostic computer-adaptive vocabulary knowledge measure: The Vocabulary Knowledge Profiler. Instead of simply following traditions and producing yet another reiteration of the conventions of established vocabulary tests, this test development project started from scratch by questioning the underlying assumptions and trying to make design decisions based not only on theoretical considerations but empirical evidence.

For this, the first step was to review the research literature and identify a) a need for a new and improved measure and b) the weaknesses of existing vocabulary tests so as to build on their strengths and attempt to overcome their flaws. The review concluded that tests of vocabulary suffer from six major weaknesses, three of which were addressed within the scope of this project so far: (1) selection of item formats, (2) sampling in terms of unit of counting, frequency bands and representativeness, and (3) the general lack of validation evidence and validation models. These issues were explored across four studies in this thesis to design a novel instrument and gather initial validation evidence for it along the way.

As is usual in test development projects, the design of this instrument and the studies it is built on highlighted the ever-present tension between theoretical ideals and practical realities. In implementing principles into practice, there are inevitable constraints that affect any test development and the (validation) research needs to focus on what is doable in order to bring about improvement in instruments step-by-step. These practical constraints range from limited sample sizes in terms of items and participants, ethics procedures that make it challenging to recruit candidatures from particular age groups in larger numbers, selecting item formats that are less than ideal in terms of score interpretation, dependencies on software programming and the funding necessary for it, workable construct definitions in between conflicting poles of single word versus multiword units, size versus depth of knowledge needs, authentic embeddedness versus construct-irrelevant variance, and partial versus precise knowledge demonstration, creating feasible test specifications that allow for systematic item development even if they imply limitations in the interpretability of scores, the conundrum of setting up a one-size-fits-all system that enables test version generation that avoids one-size-fits-all tests, to finding a balance between the measurement points needed for robust inferences and the amount of test items that test takers can be presented with without being overtaxed. "[W]e all have things to learn from relating principles to practice" (Alderson, Clapham, & Wall, 1996, p. 3) and we can often only address one challenge at a time, one study at a time, and attempt to make the best possible compromise between what is feasible and what will give us relevant and useful insights into matters yet poorly understood. This project has aimed to do just that and has, despite these tensions, generated research findings and a first version of a new measurement tool that can help inform and improve future vocabulary test development projects.

The first set of studies presented in Chapter 3 investigated the usefulness and informativeness of different item formats for vocabulary tests. Since meaning recall formats are too impractical for use on a larger scale, particularly for tests automatically scored by a machine, four different item formats were compared for how well they represented this type of form-meaning link knowledge as this is the type of knowledge that enables reading. Two separate studies with candidate groups with different characteristics were conducted to compare one form recognition, one meaning recognition and two form recall item types with a criterion meaning recall measure. It was found that all formats represented meaning recall knowledge almost equally well, but all with an unsatisfactory error in measurement of about 25%. While the formats behaved very differently individually, the MC format was identified as likely to be the most useful because of the systematic overestimation of scores associated with the item type. The findings suggested that this could be accounted for with a correction formula and the multiple-choice format was therefore selected for the development of the Vocabulary Knowledge Profiler. The studies also found a limited hierarchical link to other aspects of word knowledge, such as collocational and derivational knowledge, in terms of an implicational scale. Scores on a form-meaning link test are only partially representative of a deeper vocabulary knowledge, which should be taken into account when interpreting test scores. The results of the tests of receptive derivative knowledge administered in the study were also taken as backing for the argument of adopting the lemma as a counting and sampling unit for vocabulary tests. The assumed but unstable relationship between the knowledge of one word family member's meaning and that of other members of a word family were further problematized in the following section.

The thesis set out to identify a valid sample population of vocabulary items for a diagnostic vocabulary test based on empirical principles. Chapter 4 presented two studies, one using coverage research from large corpora and one administering a vocabulary test to proficient readers, to collect evidence on what a suitable sample population was for the present purpose. The evidence supported a cut-off at 10,000 lemmas for the Vocabulary Knowledge Profiler. The coverage findings and comparisons of word frequency lists further suggested that a new approach to frequency banding might be a useful way forward for diagnostic tests. Based on this, the Vocabulary Knowledge Profiler operationalizes a clustering of six frequency bands of 500 lemmas each for high-frequency lemmas, three 1K clusters for the mid-frequency vocabulary between the most frequent 3,000 and 6,000 lemmas and two larger clusters of 2,000 lemmas each for the two lowest frequency levels. Narrow frequency bands at the high-frequency end should thereby account for the relative significance of these bands for learners in terms of coverage, while broader bands towards the lower-frequency end of the continuum accommodate for the decrease in clustering power of the frequency factor and the fact that a lower frequency word's rank in a frequency list is increasingly dependent on the nature of the corpus.

Based on these foundation studies, test specifications were drawn up and an item bank was created, which was subjected to a large scale trial with an international candidature. Item analyses were conducted and functioning items retained for an item pool that a computer-adaptive algorithm could administer items from. It was then necessary to find out which computer-adaptive design approach would produce more reliable and more valid results.

For this purpose, a study was conducted to compare two different computer-adaptive algorithms for implementation in the test design. Chapter 6 reports on the examination of the two approaches: "floor first" and "multi-stage multi-level". Two separate studies suggested that a "floor first" design, in which candidates progress from high-frequency items to lower-frequency items until they fall below a predetermined accuracy threshold, generated more consistent scores in terms of retest-reliability. Across all frequency levels, this version demonstrated satisfactory reliability. The "floor first" version scores also showed generally higher correlations with scores from a larger and more representative item sample per frequency band in a small scale study. The design was found to provide a reasonable estimation of test takers' vocabulary knowledge at each of the frequency levels and was also argued to be more practical and potentially more motivating. The "floor first" design was therefore implemented in the Vocabulary Knowledge Profiler and the test specifications adapted accordingly. The results from these two studies also provided key elements to the draft assessment use argument for the instrument's validation.

The final study presented in this thesis attempted to collect further validation evidence by investigating the relationship between the Vocabulary Knowledge Profiler scores at different frequency bands and a test of reading comprehension. The rationale behind this study was that the score profiles of a useful test of written receptive vocabulary knowledge should be able to distinguish clearly between readers of different proficiency levels. The findings of the study were promising, but to be taken with some caution due to the limited sample population. Indeed, the frequency profiles generally followed the expected pattern. The profiles and mean % scores per frequency band were mostly in the expected order of proficiency with higher means for higher proficiency groups in almost all frequency bands. There was tentative evidence for the instrument's ability to differentiate between proficiency levels, mainly as of the fourth frequency band. However, the test did not manage to significantly distinguish between readers of adjacent proficiency levels across all bands. Most concerningly, no difference was found between scores of A2 readers and those of B1 readers. Clearly, additional data must be gathered before any robust claims about the validity of the vocabulary test can be made. Further evidence is also needed so that a key feature in the score reporting can be implemented: typical vocabulary knowledge profiles of different reader proficiency groups are to be incorporated into the score report and feedback so that test users can interpret their score meaningfully in terms of language use and compare their lexical gaps in reference to different proficiency groups.

The findings of the studies presented throughout the thesis were then pulled together to produce an initial version of a validation argument. The structure of Bachman and Palmer's (2010) Assessment Use Argument was chosen as the blueprint for the claims, warrants and backings needed for the validation of the new instrument. In doing so, it was hoped to set a model for vocabulary test validation in the future as this framework has a clear focus on the test purpose and use and it considered state-of-the-art in the field of language test validation. The argument has both incorporated evidence presented in this thesis as well as pinpointed the need for further research necessary before the launch of the test. This approach appears innovative in and of itself as most existing vocabulary tests have been published with little or no validation evidence at the time of their launch.

The validation argument therefore outlines three main points. First, it documents and describes the development of the Vocabulary Knowledge Profiler, which has tried to follow closely the criteria of useful diagnostic instruments set out by Alderson and Huhta (2011). The new test is discrete point and focuses on the linguistic element of written receptive form-meaning link knowledge. It thereby places an emphasis on language rather than language skills, and on the basic element of form-meaning link knowledge at that. It is therefore also less authentic as it tests the most frequent meaning sense of an L2 form in a minimal non-defining context. Its construct definition is informed by current SLA and vocabulary acquisition theory. It is intended for low-stakes use and takes advantage of computer-adaptive technology, for which a bespoke platform and website was designed. This means that the score report and feedback of the Vocabulary Knowledge Profiler is immediately accessible to test takers upon test completion. The feedback is detailed and meaningful through the score's links to coverage research and CEFR reading levels. This identification of lexical gaps and weaknesses is supposed to lead to and guide further learning and instruction. Finally, care has been taken to eliminate negative or inhibiting affective barriers through the computer-adaptive nature of the test and the low stakes associated with its use.

Second, the argument provides a structured overview of the validity evidence gathered to date. It clearly outlines claims about the instrument's assessment records, score interpretations, viable decisions and intended uses and consequences. Warrants and backing are provided wherever applicable and available to convince users that assessment records of the Vocabulary Knowledge Profiler are consistent across different administrations and across different test taker groups, that the interpretations about the written receptive vocabulary knowledge at the form-meaning link level assessed in the Vocabulary Knowledge Profiler are meaningful with regards to the vocabulary knowledge needed to be employed when performing reading tasks and with respect to general SLA theory of vocabulary learning, impartial to all groups of test takers, generalizable to subsequent learning activities, relevant and sufficient for the formative and diagnostic decisions to be made, that the diagnostic decisions that are made on the basis of the interpretation take into consideration existing educational and societal values and relevant laws, rules, and regulations, and are equitable for EFL learners, and that ultimately the consequences of using the Vocabulary Knowledge Profiler and of the diagnostic decisions that are made based on its results are beneficial to the EFL learners and EFL teachers using the test in their classroom learning and teaching.

And third, it clearly identifies what evidence still needs to be gathered to complete a reasonably convincing assessment use argument. In the case of the

Vocabulary Knowledge Profiler, at least four further investigations or re-runs of studies with additional data need to be conducted but were outside the scope of this PhD project. First, additional data with lower-proficiency and younger EFL learners needs to be gathered to corroborate the quality and functioning of the items in the database. The methodological procedure for this will be identical to that of the piloting described in Chapter 5, but will include younger and lower-level test takers. Ideally, around 150 responses per item will be aimed for. This will also provide a solid data basis for an extensive DIF analysis of items, which may have implications for the item pool. Second, the study presented in Chapter 7 needs to be expanded with more candidates per CEFR proficiency level, and possibly even with other proficiency tests to establish a) the typical vocabulary knowledge profiles of language users at different levels, and b) the validity of the test in its ability to distinguish clearly between the profiles of different learners. In addition to this, a concurrent validity study should be conducted in which the VKP's ability to predict reading comprehension scores should be compared with that of an existing vocabulary measure such as the VLT or the VST. If the VKP performs better in explaining reading test variance than current vocabulary tests, this would further support the validation argument. Third, a variation of the study presented in Chapter 3 should be conducted again with the finalized test, i.e. a comparison of test scores from the Vocabulary Knowledge Profiler with the vocabulary knowledge of learners as verified in a written or oral meaning recall criterion measure. It is envisaged that around 100 learners take the Vocabulary Knowledge Profiler first and then will be asked to recall the meaning of words they encountered in the test in written form. With a smaller sample, this should also be done through meaning recall interviews to confirm the results from the written meaning recall measure. Since the study in this thesis was carried out with learners from one level and one L1 background only, this needs to be investigated again with a more diverse sample population. This will then not only hopefully corroborate the meaningfulness of score interpretations, but also inform the implementation of an adjustment formula, which may be different at different frequency and proficiency levels. In the course of this, issues of test taking strategies and guessing will also have to be monitored. This will be part of the small scale interview study, but could possibly also be investigated through a variety of methodologies, such as eye-tracking, stimulated recalls or think-aloud protocols and will also allow further probing of the "I Don't Know Option" that has been problematized by researchers (Stoeckel, Bennett, & McLean, 2016; Zhang, 2013). Fourth, after completion of the user manuals, a study needs to be carried out to examine the perceived usefulness and comprehensibility of both the test, its score report and the user manuals with both EFL learners and teachers as the main stakeholders. The plan is to investigate this through a user survey tool that questions both teachers and learners about the comprehensibility, face validity, clarity and usefulness of the test, its generated profiles, and the manual. Research investigating the actual learning benefits, for instance through a classroom study on the effectiveness of tailored material design on the basis of Vocabulary Knowledge Profiles, would further enrich the validation argument. Only in this way will it be possible to provide backing for the warrants relating to the claim about the Vocabulary Knowledge Profiler's diagnostic value and its beneficial consequences for EFL vocabulary learning. With additional empirical evidence on these four issues the assessment use argument will be considered convincing enough and the test will be launched in its first version. Both test development and test validation are ongoing processes beyond what has been considered the minimal validation requirements for the present test and its purpose. The work presented in this thesis is but a first step towards an effort to providing improved diagnostic and technologically-enhanced vocabulary tests to EFL learners and teachers and challenging the field of vocabulary assessment to raise awareness for language test validation concerns.

10. References

Abdullah, K. I., Puteh, F., Azizan, A. R., Hamdan, N. N., & Saude, S. (2013). Validation of a controlled productive Vocabulary Levels Test below the 2000-word level. *System*, *41*(2), 352–364. http://doi.org/10.1016/j.system.2013.03.005

Abels, M. (1994). *Ken ik dit woord?* Unpublished PhD thesis. Catholic University of Nijmegen.

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438.

Aitchison, J. (2003). *Words in the mind*. Oxford: Blackwell.

Aizawa, K. (2006). Rethinking frequency markers for English-Japanese dictionaries. In M. Murata, K. Minamide, Y. Tono, & S. Ishikawa (Eds.), *English lexicography in Japan* (pp. 108–119). Tokyo: Taishukan Publishing Company.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, *13*, 219–227.

Alderson, J. C. (1990). Learner-centered testing through computers: Institutional issues in individual assessment. In J. de Jong & D. K. Stevenson (Eds.), *Individualizing the assessment of language abilities* (pp. 20–27). Clevedon: Multilingual Matters.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 383–409.

Alderson, J. C., Brunfaut, T., & Harding, L. (2014). Towards a Theory of Diagnosis in Second and Foreign Language Assessment: Insights from Professional Practice Across Diverse Fields. *Applied Linguistics*, 1–26. http://doi.org/10.1093/applin/amt046

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., & Cseresznyés, M. (2003). *Into Europe – Prepare for Modern English Exams: Reading and Use of English.* Teleki Lazlo Foundation and The British Council Hungary.

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The Diagnosis of Reading in a Second or Foreign Language*. New York: Routledge.

Alderson, J. C., & Huhta, A. (2011). Can research into the diagnostic testing of reading in a second or foreign language contribute to SLA research? In L. Roberts, G. Pallotti, & C. Bettoni (Eds.), *EUROSLA Yearbook 11* (Amsterdam, pp. 30–52). John Benjamins.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson (Ed.), *Advances in reading/language research: A research annual* (pp. 231–256). Greenwich, CT: JAI Press.

Arnaud, P. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and Problems in Language Testing* (pp. 14–28). Colchester: University of Essex.

Arnaud, P. (1989). Vocabulary and grammar: a multitrait-multimethod investigation. *AILA Review*, *6*, 56–65.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*, 535–556.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: OUP Oxford.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: OUP Oxford.

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford: OUP Oxford.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System*, *21*, 101–114.

Barclay, S. (2013). A test of aural vocabulary for Japanese students. *Vocabulary Education & Research Bulletin*, *2*(2), 7–8.

Barfield, A. (2003). *Collocation recognition and production: Research insights*. Chuo University Japan.

Barnett, M. A. (1986). Syntactic and lexical/semantic skills in foreign language reading: importance and interaction. *Modern Language Journal1*, *70*, 343–349.

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, *27*(2), 223–247. http://doi.org/10.1016/S0346-251X(99)00018-4

Bauer, L., & Nation, I. S. P. (1993). Word Families. *International Journal of Lexicography*, *6*(4), 253–279.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: some methodological issues in theory and practice. *Language Testing*, *18*(3), 235–274. http://doi.org/10.1177/026553220101800301

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

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing*, *16*(2), 131–162. http://doi.org/10.1177/026553229901600202

Beks, B. (2001). *Le degré des connaissances lexicales [The degree of lexical knowledge]*. Amsterdam: Vrije Universitet Amsterdam.

Bertram, R., Baayen, R., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, *42*, 390–405.

Bertram, R., Laine, M., & Virkkala, M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology*, *41*(4), 287–296.

Bogaards, P. (2000). Testing L2 vocabulary knowledge at a high level: The case of the Euralex French tests. *Applied Linguistics*, *21*(4), 490–516.

Bonk, W. J. (2001). Testing ESL learners' knowledge of collocations. In T. D. Hudson & J. D. Brown (Eds.), *A Focus on Language Test Development: Expanding the Language Proficiency Construct Across a Variety of Tests. (Technical Report #21)* (pp. 113–142). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Brezina, V., & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*, *36*(1), 1–22. http://doi.org/10.1093/applin/amt018

Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 171–174.

Brisbois, J. (1995). Connections between first- and second-language reading. *Journal of Literacy Research*, *27*(4), 565–584. http://doi.org/10.1080/10862969509547899

British Council. (2013). Aptis Candidate Guide. Retrieved from https://www.britishcouncil.org/sites/default/files/Aptis-candidate-guide-web.pdf

Brown, D. (2012). The frequency model of vocabulary learning and Japanese learners. *Vocabulary Learning and Instruction*, *1*(1), 20–28.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. London.

Bruton, A. (2009). The Vocabulary Knowledge Scale: A critical analysis. *Language Assessment Quarterly*, *6*(4), 288–297.

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, *6*(2), 145–173. http://doi.org/10.1191/1362168802lr103oa

Chapelle, C. (1994). Are C-Tests valid measures for L2 vocabulary research? *Second Language Research*, *10*, 157–187.

Chapelle, C. (2012). Validity argument for language assessment: The framework is simple... *Language Testing*, *29*(1), 19–27.

http://doi.org/10.1177/0265532211417211

Chapelle, C., & Douglas, D. (2006). *Assessing Language through Computer Technology*. Cambridge: Cambridge University Press.

Coady, J., Magoto, J., Hubbard, P., Graney, J., & Mokhtari, K. (1993). High frequency vocabulary and reading proficiency in ESL readers. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second language reading and vocabulary learning* (pp. 217–228). Norwood, NJ: Ablex.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System*, *25*, 201–315.

Cobb, T. (2000). One size fits all? Francophone learners and English vocabulary tests. *Canadian Modern Language Review*, *57*(2), 295–324.

Corrigan, A., & Upshur, J. A. (1982). Test method and linguistic factors in foreign language tests. *IRAL*, *20*, 313–321.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2014). *Linguistic Integration of Adult Migrants: Policy and practice - Final Report on the 3rd Council of Europe Survey*. Strasbourg.

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

Coxhead, A., Nation, P., & Sim, D. (2015). Vocabulary size and native speaker secondary school students. *New Zealand Journal of Educational Studies*. http://doi.org/10.1007/s40841-015-0002-3

Cremer, M., Dingshoff, D., de Beer, M., & Schoonen, R. (2010). Do word associations assess word knowledge? A comparison of L1 and L2, child and adult word associations. *International Journal of Bilingualism*, *15*(2), 187–204.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner tests using computational indices. *Language Testing*, *28*(4), 561–580.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press.

Daller, H., & Phelan, D. (2007). What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 234–244). Cambridge: Cambridge University Press.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197–222.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 150–164). Cambridge: Cambridge University Press.

Dang, T. N. Y., & Webb, S. (2016). Making an essential word list for beginners. In I. S. P. Nation (Ed.), *Making and using word lists for language learning and testing* (pp. 153–167). Amsterdam: John Benjamins.

Davidson, F. (2012). Test Specifications. In *The Encyclopedia of Applied Linguistics*.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven: Yale University Press.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 520 million words, 1990-present. Retrieved from http://corpus.byu.edu/coca/.

de la Fuente, M. J. (2002). Negotiation and oral acquisition of vocabulary. *Studies in Second Language Acquisition*, *24*, 81–112.

Dewaele, J.-M., & MacIntyre, P. D. (2014). The two faces of Janus. Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, *4*(2), 237–274.

Dolch, E. W., & Leeds, D. (1953). Vocabulary tests and depth of meaning. *Journal of Educational Research*, *4*, 181–189.

Dörnyei, Z. (2001). *Motivational strategies in the language classroom*. Cambridge: Cambridge University Press.

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 9–42). Bristol, UK: Multilingual Matters.

Dörnyei, Z. (2014). Motivation in second language learning. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language (4th ed.)* (pp. 518–531). Boston, MA: National Geographic Learning/Cengage Learning.

Dörnyei, Z., Ibrahim, Z., & Muir, C. (2015). "Directed Motivational Currents": Regulating complex dynamic systems through motivational surges. In Z. Dörnyei, P. D. MacIntyre, & A. Henry (Eds.), *Motivational dynamics in language learning* (pp. 95–105). Bristol, UK: Multilingual Matters.

Dörnyei, Z., Muir, C., & Ibrahim, Z. (2014). Directed Motivational Currents: Energising language learning through creating intense motivational pathways. In D. Lasagabaster, A. Doiz, & J. M. Sierra (Eds.), *Motivation and foreign language learning: From theory to practice* (pp. 9–29). Amsterdam: John Benjamins.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. New York: Routledge.

Dörnyei, Z., & Ushioda, E. (2011). *Teaching and researching motivation*. Harlow: Longman.

Dunn, M., & Dunn, L. M. (1959). Peabody Picture Vocabulary Test. Circle Pines, MN: AGS.

Dunn, M., & Dunn, L. M. (2007). Peabody Picture Vocabulary Test-4. Circle Pines, MN: AGS.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325.

Edwards, R., & Collins, L. (2011). Lexical frequency profiles and Zipf's Law. *Language Learning*, *61*(1), 1–30.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*(2), 253–272. http://doi.org/10.1177/0265532212459028

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62.

Eyckmans, J. (2004). *Measuring receptive vocabulary size*. LOT (Landelijke Onderzoekschool Taalwetenschap).

Eyckmans, J. (2009). Towards 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). New York: Palgrave Macmillan.

Eyckmans, J., Van de Velde, H., van Hout, R., & Boers, F. (2007). Learners' response behaviour in Yes/No Vocabulary Tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 59–76). Cambridge: Cambridge University Press.

Farghal, M., & Obiedat, H. (1995). Collocations: A neglected variable in EFL. *International Journal of Applied Linguistics*, *28*(4), 313–331.

Fitzpatrick, T. (2007). Productive vocabulary tests and the search for concurrent validity. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 116–132). Cambridge: Cambridge University Press.

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, *27*(4), 537–554. http://doi.org/10.1177/0265532209354771

Fitzpatrick, T., & Clenton, J. (2017). Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge. *TESOL Quarterly*, *Online fir*, 1–24.

Fitzpatrick, T., & Meara, P. (2004). Exploring the validity of a test of productive vocabulary. *Vigo International Journal of Applied Linguistics*, *1*, 55–74.

Fountain, R. L., & Nation, I. S. P. (2000). A vocabulary-based graded dictation test. *RELC Journal*, *31*(2), 29–44.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, *20*, 384–408.

Fulcher, G. (2014). Philosophy and Language Testing. In A. J. Kunnan (Ed.), *The Companion to Language Testing* (pp. 1431–1451). London: Wiley-Blackwell.

Fulcher, G. (2015). *Re-examining Language Testing: a philosophical and social enquiry*. London & New York: Routledge.

Geranpayeh, A., & Taylor, L. (Eds.). (2013). *Examining listening: Research and practice in assessing second language listening.* Cambridge: Cambridge University Press.

Golonka, E., Bowles, A., Silbert, N., Kramasz, D., Blake, C., & Buckwalter, T. (2015). The Role of Context and Cognitive Effort in Vocabulary Learning: A Study of Intermediate-Level Learners of Arabic. *Modern Language Journal*, *99*(1), 19–39.

González-Fernández, B., & Schmitt, N. (under review). Order of acquisition and relationships among word knowledge components.

Goodrich, H. C. (1977). Distractor efficiency in foreign language testing. *TESOL Quarterly*, *11*(1), 69–78.

Goulden, R., Nation, I. S. P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*(4), 341–363.

Green, R. (2013). *Statistical Analyses for Language Testers*. Basingstoke: Palgrave Macmillan.

Greidanus, T., Bogaards, P., van der Linden, E., Nienhuis, L., & de Wolf, T. (2004). The construction and validation of a depp word knowledge test for advanced learners of French. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 191–208). Amsterdam: John Benjamins.

Greidanus, T., & Nienhuis, L. (2001). Testing the Quality of Word Knowledge in a Second Language by Means of Word Associations : Types of Distractors and Types of Associations. *Modern Language Journal*, *85*(5), 567–577.

Gyllstad, H. (2005). *Words that go together well: developing test formats for measuring learner knowledge of English collocations*. The Department of English in Lund: Working Papers in Linguistics, Volume 5.

Gyllstad, H. (2007). *Testing English Collocations*. Unpublished PhD thesis. University of Lund, Sweden.

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). New York: Palgrave Macmillan.

Gyllstad, H. (2012). Validating the Vocabulary Size Test: A classical test theory approach. Poster presented at the 9th annual EALTA conference. University of Innsbruck, Austria.

Gyllstad, H., Vilkaité, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics*, *166*(2), 278–306.

Hacquebord, H. (1999). Less- en luisterbegrip van studieteksten bij Nederlandse en anderstalige leerlingen en studenten. In E. Huls & B. Weltens (Eds.), *Artikelen van de Derde Sociolinguistische Conferentie*. Delft: Eburon.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: elaborating on diagnostic principles. *Language Testing*, 1–20.

Henning, G. (1987). *A guide to language testing*. Cambridge, Mass.: Newbury House.

Henning, G. H. (1975). Measuring foreign language reading comprehension. *Language Learning*, *25*, 109–114.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching Collocation: Further developments in the Lexical Approach* (pp. 47–69). Hove: LTP.

Hoffman, L., Templin, J., & Rice, M. L. (2012). Linking Outcomes from Peabody Picture Vocabulary Test Forms Using Item Response Models. *Journal of Speech, Language and Hearing Research*, *55*, 754–763.

Holster, T. E., & Lake, J. (2016). Guessing and the Rasch Model. *Language Assessment Quarterly*, *13*(2), 124–141.

Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with an interactive on-line database. *Language Learning and Technology*, *9*(2), 99–110.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Hughes, A. (2003). *Testing for Language Teachers (Cambridge Language Teaching Library)*. Cambridge: Cambridge University Press ELT.

Huhta, A., Alderson, J. C., Nieminen, L., & Ullakonoja, R. (2011). Diagnosing reading in L2 – predictors and vocabulary profiles. Paper presented at ACTFL CEFR Alignment Conference 2011. Provo.

Huibregtse, I., & Admiraal, W. (1999). De score op een ja/nee woordenschattoets: correctie voor raden en persoonlijke antwoordstijl. *Tijdschrift Voor Onderwijsresearch*, *24*, 110–124.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes–no vocabulary test: correction for guessing and response style. *Language Testing*, *19*(3), 227–245. http://doi.org/10.1191/0265532202lt229oa

Hulstijn, J. H., Schoonen, R., De Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203–221.

In'nami, Y. (2013). Second-Language Vocabulary Assessment Research: Issues and Challenges. *Vocabulary Learning and Instruction*, *2*(1), 74–81.

Ishii, T., & Schmitt, N. (2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal*, *40*(5), 5–22.

JACET. (2003). *JACET list of 8000 basic words.* Tokyo: Japan Association of College English Teachers.

Jeon, E. H., & Yamashita, J. (2014). L2 Reading Comprehension and Its Correlates: A Meta-Analysis. *Language Learning*, *64*(1), 160–212.

Jonz, J. (1990). Another turn in the conversation: what does cloze measure? *TESOL Quarterly*, *24*, 61–83.

Jun Zhang, L., & Bin Anual, S. (2008). The Role of Vocabulary in Reading Comprehension: The Case of Secondary School Students Learning English in Singapore. *RELC Journal*, *39*(1), 51–76. http://doi.org/10.1177/0033688208091140

Kamimoto, T. (2008). Guessing and vocabulary tests: Looking at the Vocabulary Levels Test. Paper presented at the 41 Annual BAAL Conference held in Swansea, 11-13 September 2008.

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

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 135–170.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17–64). Westport, CT: American Council on Education/Praeger Publishers.

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17. http://doi.org/10.1177/0265532211417210

Karami, H. (2012). The Development and Validation of a Bilingual Version of the Vocabulary Size Test. *RELC Journal*, *43*(1), 53–67.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, *68*(8), 1665–92.

Khalifa, H., & Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. Cambridge: Cambridge University Press.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., … Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1–30.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-Test. *Language Testing*, *1*, 134–146.

Koda, K. (1989). The effects of transferred vocabulary knowledge on the development of L2 reading proficiency. *Foreign Language Annals*, *22*, 529–540.

Koda, K. (2004). *Insights into second language reading: a cross-linguistic approach*. Cambridge: Cambridge University Press.

Kramer, B., & McLean, S. (2015). Comparing aural and written receptive vocabulary knowledge of the first 5k and the AWL. Paper presented at

LTRC 2015, Toronto, Canada.

Kremmel, B. (2016). Word Families and Frequency Bands in Vocabulary Tests: Challenging Conventions. *TESOL Quarterly*, *50*(4), 976–987.

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2015). Exploring the Role of Phraseological Knowledge in Foreign Language Reading. *Applied Linguistics*, *Advance Access*, 1–24.

Kremmel, B., & Schmitt, N. (2016). Interpreting Vocabulary Test Scores: What Do Various Item Formats Tell Us about Learners' Ability to Employ Words? *Language Assessment Quarterly*, *13*(4), 377–392.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. London: Longman.

Laufer, B. (1992a). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). London: Macmillan.

Laufer, B. (1992b). Reading in a foreign language: how does L2 lexical knowledge interact with the reader's general academic ability? *Journal of Research in Reading*, *15*, 95–103.

Laufer, B. (1998). The Development of Passive and Active Vocabulary in a Second Language : Same or Different ? *Applied Linguistics*, *12*, 255–271.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226. http://doi.org/10.1191/0265532204lt277oa

Laufer, B., & Goldstein, Z. (2004). Testing Vocabulary Knowledge : Size, Strength , and Computer Adaptiveness. *Language Learning*, *54*(3), 399–436.

Laufer, B., & McLean, S. (2016). Loanwords and Vocabulary Size Test Scores: A Case of Different Estimates for Different L1 Learners. *Language Assessment Quarterly*, *13*(3), 202–217.

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

Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. http://doi.org/10.1177/026553229901600103

Laufer, B., & Nation, I. S. P. (2001). Passive vocabulary size and speed of meaning recognition. In *EUROSLA Yearbook 1* (pp. 7–28).

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: effects of language learning context. *Language Learning*, *48*, 365–391.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

Leech, G., Deuchar, M., & Hoogenraad, R. (1982). *English Grammar For Today*. London: Macmillan.

Lemhoefer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*, 325–343.

Levitzky-Aviad, T., Mizrahi, E., & Laufer, B. (2014). A new test of active vocabulary size. In *EUROSLA 24 book of abstracts* (p. 48).

Lewis, M. (1993). *The Lexical Approach: The State of ELT and a Way Forward*. Hove: Language Teaching Publications.

Linacre, J. M. (2017). Winsteps® Rasch measurement computer program. Beaverton, Oregon. Retrieved from winsteps.com

Long, M. H., & Richards, J. C. (2007). Series Editor's Preface. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. xii–xiii). Cambridge: Cambridge University Press.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet-assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*, 189–202.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229–249.

Macis, M. (2013). Partners vs. Phrasemes: A new conceptualisation of collocation. Poster presented at Language Testing Forum 2013, University of Nottingham, UK.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language: Papers from the annual meeting of the BAAL held at the University of Wales, Swansea, September 1996* (pp. 58–71). Clevedon: Multilingual Matters.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave Macmillan.

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics*, *33*(3), 299–320. http://doi.org/10.1093/applin/ams010

McLean, S. (2017). Evidence for the Adoption of the Flemma as an Appropriate Word Counting Unit. *Applied Linguistics Advance access, p. 1-24*.

McLean, S., Kramer, B., & Beglar, D. (2015). The Creation and Validation of a Listening Vocabulary Levels Test. *Language Teaching Research*, *19*(6), 741–760.

McNamara, T. (2000). *Language Testing*. Oxford: OUP Oxford.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31–51.

McNeill, A. (1996). Vocabulary knowledge Profiles: Evidence from Chinese-speaking ESL teachers. *Hong Kong Journal of Applied Linguistics*, *1*, 39–63.

Meara, P. (1992). *EFL Vocabulary Tests*. Swansea: Lognostics.

Meara, P. (1994). The complexities of simple vocabulary tests. In F. G. Brinkman, J. A. van der Schee, & M. C. Schouten-vanParreren (Eds.), *Curriculum research: different disciplines and common goals* (pp. 15–28). Amsterdam: Vrije Universiteit.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and Competence in Second Langauge Acquisition* (pp. 35–53). Cambridge: Cambridge University Press.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing lexical characteristics of short texts. *Prospect*, *16*(3), 5–19.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–154.

Meara, P., & Fitzpatrick, T. (2000). Lex30: an improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19–30. http://doi.org/10.1016/S0346-251X(99)00058-5

Meara, P., & Jones, G. (1990). Eurocentres Vocabulary Size Test, Version E1.1/K10. Zurich: Eurocentres Learning Service.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy* (pp. 84–102). Cambridge: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13–104). New York: Macmillan.

Miller, G. A. (1999). On knowing a word. *Annual Review of Psychology*, *50*(1), 1–19.

Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 47–58). Cambridge: Cambridge University Press.

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Bristol, UK: Multilingual Matters.

Milton, J., & Hopkins, N. (2005). A_Lex. Swansea: University of Wales Swansea.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review*, *63*(1), 127–147.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacon-Beltran, C. Abello-Contesse, M. Torreblanca-Lopez, & M. D. Lopez-Jimenez (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83–97). Bristol: Multilingual Matters.

Mislevy, R. J. (2007). Validity by Design. *Educational Researcher*, *36*(8), 463–469. http://doi.org/10.3102/0013189X07311660

Mizumoto, A., Sasao, Y., & Webb, S. (2017). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test.

Language Testing, *Online fir*, 1–23.

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73–98. http://doi.org/10.1191/0265532206lt321oa

Mochizuki, M. (2002). Exploration of two aspects of vocabulary knowledge: paradigmatic and collocational. *Annual Review of English Language Education in Japan*, *13*, 121–129.

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy* (pp. 40–63). Cambridge: Cambridge University Press.

Nagy, W. E., Anderson, R., Schommer, M., Scott, J. A., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, *24*(3), 263–282.

Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *Modern Language Journal*, *87*, 261–276.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. Heinle and Heinle.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

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. 3–13). Amsterdam: John Benjamins.

Nation, I. S. P. (2006). How Large a Vocabulary Is Needed For Reading and Listening ? *Canadian Modern Language Review*, *63*(1), 59–82.

Nation, I. S. P. (2008). *Teaching Vocabulary - Strategies & Techniques*. Boston, MA: Heinle-Cengage ELT.

Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching*, *44*(4), 529–539. http://doi.org/10.1017/S0261444811000267

Nation, I. S. P. (2014). Vocabulary Size Tests. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

Nation, I. S. P., & Anthony, L. (2016). The Picture Vocabulary Size Test. Poster presentation given at Vocab@Tokyo 2016, Meiji Gakuin University, Tokyo, Japan.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nation, I. S. P., & Gu, P. Y. (2007). *Focus on Vocabulary*. Sydney: NCELTR Publications.

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*. Cambridge: Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). *Researching Vocabulary*. Boston, MA: Heinle-Cengage ELT.

Nguyen, L., & Nation, I. S. P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99.

North, B. (2004). Relating assessments, examinations, and courses to the CEF. In K. Morrow (Ed.), *Insights from the Common European Framework* (pp. 77–90). Oxford: Oxford University Press.

Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, *18*, 161–175.

O'Sullivan, B. (2015). *Technical Report: Aptis Test Development Approach*. Retrieved from http://www.britishcouncil.org/sites/britishcouncil.uk2/files/tech_001_barry_osullivan_aptis_test_-_v5.pdf

O'Sullivan, B., & Dunlea, J. (2015). *Aptis general technical manual*.

Paek, P. (2005). *Recent trends in comparability studies*. Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/TrendsCompStudies.pdf?%0AWT.mc_id=TMRS_Recent_Trends_in_Comparability_Studies

Paribakht, T. S. (2005). The influence of L1 lexicalization of L2 lexical inferencing: A study of Farsi-speaking EFL learners. *Language Learning*, *55*(4), 701–748.

Paribakht, T. S., & Wesche, M. B. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition* (pp. 174–200). Cambridge: Cambridge University Press.

Parry, K. J. (1987). Reading in a second culture. In J. Devine, P. L. Carrell, & D. E. Eskey (Eds.), *Research in reading in English as a second language* (pp. 59–70). Washington, DC: TESOL.

Paul, P. V., Stallman, A., & O'Rourke, J. P. (1990). *Using three test formats to assess good and poor reader's word knowledge. Technical Report No. 509 of the Center for the Study of Reading*.

Pearson. (2014). *PPVT-4 Technical Specifications*. Retrieved from http://images.pearsonclinical.com/images/Products/PPVT-IV/ppvt4.pdf

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes-No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489–509. http://doi.org/10.1177/0265532212438053

Petrescu, M. C., Helms-Park, R., & Dronjic, V. (2017). The impact of frequency and register on cognate facilitation: Comparing Romanian and Vietnamese speakers on the Vocabulary Levels Test. *English for Specific Purposes*, *47*, 15–25.

Pike, L. W. (1979). *An evaluation of alternative item formats for testing English as a foreign language (TOEFL Research Report 2)*. Princeton, NJ: Educational Testing Service.

Pinchbeck, G. (2016). Developmental scales of L1 and L2 academic English vocabulary: Vocabulary test item difficulty indicates lexical sophistication and derivational morphology development. Paper presented at the American Association for Applied Linguistics conference, Orland.

Porter, D. (1983). The effect of quantity of context on the ability to make linguistic predictions: a flaw in a measure of general proficiency. In A. Hughes & D. Porter (Eds.), *Current Developments in Language Testing* (pp. 63–74). New York: Academic Press.

Pulido, D. (2004). The effect of cultural familiarity on incidental vocabulary acquisition through reading. *The Reading Matrix*, *4*(2), 20–53.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282–308.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536.

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28–52. http://doi.org/10.1191/0265532204lt273oa

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, *19*(2), 12–25.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371.

Read, J. (1995). Refining the word associates format as a measure of depth of vocabulary knowledge. *New Zealand Studies in Applied Linguistics*, *1*, 1–17.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Laurence Erlbaum.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2004a). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 209–227). Amsterdam: John Benjamins.

Read, J. (2004b). Second Language Vocabulary Testing : Taking a Broader Perspective. *Paper Presented at the 2004 Inernational Conference on English Instruction and Assessment*.

Read, J. (2007). Second Language Vocabulary Assessment: Current Practices and New Directions. *International Journal of English Studies*, *7*(2), 105–

125.

Read, J. (2013). Second language vocabulary assessment. *Language Teaching*, *46*(1), 41–52. http://doi.org/10.1017/S0261444812000377

Read, J., & Chapelle, C. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32. http://doi.org/10.1177/026553220101800101

Revier, R. L. (2009). Evaluating a New Test of Whole English Collocations. In A. Barfield & H. Gyllstad (Eds.), *Researching Collocations in another language: Multiple interpretations* (pp. 125–138). Palgrave Macmillan.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*(1), 77–89.

Rocca, L. (2017). A perspective on language testing in the context of migration and integration. Paper presented at the 14th Annual Conference of EALTA, Sèvres.

Römer, U. (2009). The inseparability of lexis and grammar<BR> Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*(1), 140–162. http://doi.org/10.1075/arcl.7.06rom

Rott, S., Williams, J., & Cameron, R. (2002). The effect of multiple-choice L1 glosses and input-output cycles on lexical acquisition and retention. *Language Teaching Research*, *6*, 183–222.

Schmitt, N. (1995). A fresh approach to vocabulary using a word knowledge framework. *RELC Journal*, *26*(1), 86–94.

Schmitt, N. (1998a). Measuring collocational knowledge: key issues and an experimental assessment procedure. *I.T.L. Review of Applied Linguistics 119-120*, 27–47.

Schmitt, N. (1998b). Tracking the Incremental Acquisition of Second Language Vocabulary : A Longitudinal Study. *Language Learning*, *48*(2), 281–317.

Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing*, *16*(2), 189–216. http://doi.org/10.1177/026553229901600204

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N. (2014). Size and Depth of Vocabulary Knowledge: What the Research Shows. *Language Learning*, *64*(4), 913–951.

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2016). How much vocabulary is needed to use English? Replication of Van Zeeland & Schmitt (2012), Nation, (2006), and Cobb (2007). *Language Teaching*.

Schmitt, N., Dörnyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use* (pp. 55–86). Amsterdam: John Benjamins.

Schmitt, N., Gardner, D., & Davies, M. (under review). How Much Vocabulary is Required for Listening and Reading in English?

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26–43.

Schmitt, N., & McCarthy, M. (1997). Introduction. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy* (p. 1–??). Cambridge: Cambridge University Press.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*, 17–36.

Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The Word Associates Format: Validation evidence. *Language Testing*, *28*(1), 105–126. http://doi.org/10.1177/0265532210373605

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503.

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. http://doi.org/10.1177/026553220101800103

Schmitt, N., & Zimmerman, C. B. (2002). Derivative Word Forms: What Do Learners Know? *TESOL Quarterly*, *36*(2), 145. http://doi.org/10.2307/3588328

Schonell, F. J., Meddleton, I. G., & Shaw, B. A. (1956). *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.

Schoonen, R., Hulstijn, J., & Bossers, B. (1998). Metacognitive and language-specific knowledge in native and foreign language reading comprehension: an empirical study among Dutch students in grades 6, 8 and 10. *Language Learning 48 (1), 48*(1), 71–106.

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*(2), 211–236. http://doi.org/10.1177/0265532207086782

Schultz, K. S., Whitney, D. J., & Zickar, M. J. (Eds.). (2014). *Measurement theory in action (2nd Ed.)*. Hove: Taylor & Francis.

Segalowitz, N. (2015). Fluency of vocabulary access and the assessment of L2 vocabulary.

Shaw, S. D., & Weir, C. J. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing*. Cambridge: Cambridge University Press.

Shillaw, J. (1999). *The application of the Rasch model to Yes/No vocabulary tests*. University of Wales Swansea.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99–128.

http://doi.org/10.1177/0265532207071513

Silva, R., & Clahsen, H. (2008). Morphologically complex words in L1 and L2 processing: Evidence from masked priming experiments in English. *Bilingualism: Language and Cognition*, *11*(2), 245–260.

Sims, V. M. (1929). The reliability and validity of four types of vocabulary test. *Journal of Educational Research*, *20*, 91–96.

Simsek, E., & Dörnyei, Z. (2017). Anxiety and L2 self-images: The "anxious self." In C. Gkonou, M. Daubney, & J.-M. Dewaele (Eds.), *New insights into language anxiety: Theory, research and educational implications* (pp. 51–69). Bristol, UK: Multilingual Matters.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: OUP.

Sinclair, J. (2004). *Trust the Text: Language, Corpus and Discourse*. New York: Routledge.

Singleton, D. (1999). *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.

Siyanova-Chanturia, A., & Martinez, R. (2014). The Idiom Principle Revisited. *Applied Linguistics*, 1–22. http://doi.org/10.1093/applin/amt054

Snellings, P., van Gelderen, A., & de Glopper, K. (2004). Validating a test of second language written lexical retrieval: a new measure of fluency in written language production. *Language Testing*, *21*(2), 174–201. http://doi.org/10.1191/0265532204lt276oa

Sorell, C. J. (2013). *A study of issues and techniques for creating core vocabulary lists for English as an international language.* Unpublished PhD thesis. Victoria University of Wellington, New Zealand.

Sparks, R. L., Artzer, M., Ganschow, L., Siebenhar, D., Plageman, M., & Patton, J. (1998). Differences in native-language skills, foreign-language aptitude, and foreign-language grades among high-, average-, and low-proficiency foreign-language learners: two studies. *Language Testing*, *15*(2), 181–216. http://doi.org/10.1177/026553229801500203

Staehr, L. (2009). Vocabulary knowledge and advanced listening comprehension in English as a Foreign Language. *Studies in Second Language Acquisition*, *31*, 1–31.

Stalnaker, J. M., & Kurath, W. (1935). A comparison of two types of foreign language vocabulary test. *Journal of Educational Psychology*, *26*, 435–442.

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 89–103). Hillsdale, NJ: Erlbaum.

Stewart, J. (2012). A Multiple-Choice Test of Active Vocabulary Knowledge. *Vocabulary Learning and Instruction*, *1*(1), 53–69.

Stewart, J. (2014). Do Multiple-Choice Options Inflate Estimates of Vocabulary Size on the VST? *Language Assessment Quarterly*, *11*(3), 271–282.

Stewart, J., Batty, A. O., & Bovee, N. (2012). Comparing multidimensional and continuum models of vocabulary acquisition: An empirical examination of the Vocabulary Knowledge Scale. *TESOL Quarterly*, *46*(4), 695–721.

Stewart, J., McLean, S., & Kramer, B. (2017). A Response to Holster and Lake Regarding Guessing and the Rasch Model. *Language Assessment Quarterly*, *14*(1), 69–74.

Stewart, J., & White, D. (2011). Estimating Guessing Effects on the Vocabulary Levels Test for Differing Degrees of Word Knowledge. *TESOL Quarterly*, *45*(2), 370–380. http://doi.org/10.5054/tq.2011.254523

Stoeckel, T., Bennett, P., & McLean, S. (2016). Is "I Don't Know" a Viable Answer Choice on the Vocabulary Size Test? *TESOL Quarterly*, *50*(4), 965–975.

Stubbe, R. (2013). Comparing Regression versus Correction Formula Predictions of Passive Recall Test Scores from Yes-No Test Results. *Vocabulary Learning and Instruction*, *2*(1), 39–46.

Stubbe, R., & Stewart, J. (2012). Optimizing scoring formulas for yes/no vocabulary checklists using linear models. *Shiken Research Bulletin*, *16*(2), 2–7. Retrieved from http://teval.jalt.org/node/12

Stubbe, R., Stewart, J., & Pritchard, T. (2010). Examining the effects of pseudowords in yes/no vocabulary tests for low level learners. *Kyushu Sangyo University Language Education and Research Center Journal*, *5*, 5–23.

Stubbs, M. (2009). Memorial Article: John Sinclair (1933-2007): The Search for Units of Meaning: Sinclair on Empirical Semantics. *Applied Linguistics*, *30*(1), 115–137. http://doi.org/10.1093/applin/amn052

Swain, M. (1983). Large-scale communicative language testing: A case study. *Language Learning and Communication*, *2*, 133–147.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. *Language Testing*, *17*(3), 323–340. http://doi.org/10.1177/026553220001700303

Taylor, L. (Ed.). (2011). *Examining Speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.

Thekes, I. J. (2013). The correlations between the processing of English as a second language idioms, effectance motivation, native language idiom knowledge, inductive reasoning and metacognitive awareness.

Thissen, D. (2000). Reliability and Measurement Precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 159–184). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Thornbury, S. (2002). *How to Teach Vocabulary*. London: Longman.

Tilley, H. C. (1936). A technique for determining ther relative difficulty of word meanings among elementary school children. *Journal of Experimental Education*, *5*, 61–64.

Tomiyama, M. (2008). Age and Proficiency in L2 Attrition: Data from Two Siblings. *Applied Linguistics*, *30*(2), 253–275.

http://doi.org/10.1093/applin/amn038

Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Treffers-Daller, J., Parslow, P., & Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*. http://doi.org/10.1093/applin/amw009

Tseng, W. (2013). Validating a Pictorial Vocabulary Size Test via the 3PL-IRT Model. *Vocabulary Learning and Instruction*, *2*(1), 64–73.

Tseng, W. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education*, *97*, 69–85. http://doi.org/10.1016/j.compedu.2016.02.018.

Tung, P. (1986). Computer adaptive testing: Implications for language test developers. In C. Stansfield (Ed.), *Technology and language testing* (pp. 13–28). Washington, D.C.: TESOL Publications.

Unsworth, S., Persson, L., Prins, T., & De Bot, K. (2014). An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics*, 1–23. http://doi.org/10.1093/applin/amt052

van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). Roles of linguistic knowledge, metacognitive knowledge and processing speed in L3, L2 and L1 reading comprehension: a structural equation modelling approach. *International Journal of Bilingualism*, *7*(1), 7–25.

van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed and metacognitive knowledge in first and second language reading comprehension: a componential analysis. *Journal of Educational Psychology*, *96*(1), 19–30.

van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 93–115). Cambridge: Cambridge University Press.

van Zeeland, H. (2013). *Second Language Vocabulary Knowledge in and from Listening* (Unpublishe). University of Nottingham, Nottingham.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*, 457–479.

Verhallen, M., Oezdemir, L., Yueksel, E., & Schoonen, R. (1999). Woordkennis van Turkse kinderen in de bovenbouw van het basisonderwijs [Lexical knowledge of Turkish children in the upper grades of primary education]. *Toegepaste Taalwetenschap in Artikelen*, *61*, 21–33.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*, 217–234.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 173–189). Amsterdam: John Benjamins.

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*. Iowa State University.

Walters, J. (2012). Aspects of Validity of a Test of Productive Vocabulary: Lex30. *Language Assessment Quarterly*, *9*(2), 172–185.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Education and Psychological Measurement*, *68*, 5–24.

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System*, *37*(3), 461–469. http://doi.org/10.1016/j.system.2009.01.004

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*, 130–163.

Webb, N. L. (2006). Identifying Content for Student Achievement Tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 155–180). New Jersey: Lawrence Erlbaum Associates.

Webb, S. (2005). Receptive and productive voabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*, 33–52.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79–95.

Webb, S., & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335–366.

Webb, S., & Sasao, Y. (2013). New Directions In Vocabulary Testing. *RELC Journal*, *44*(3), 263–277. http://doi.org/10.1177/0033688213500582

Weir, C. J. (1990). *Communicative language testing*. London: Prentice Hall International (UK) Ltd.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53*, 774–789. http://doi.org/10.1037/0022-006X.53.6.774

Wesche, M. B., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review*, *53*(1), 13–40.

Wesche, M. B., & Paribakht, T. S. (2009). *Lexical Inferencing in a First and Second Language: Cross-linguistic Dimensions*. Clevedon: Multilingual

Matters.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of Vocabulary Levels Tests. *System*, *35*(2), 182–191. http://doi.org/10.1016/j.system.2006.12.009

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.

Yamashita, J. (1999). *Reading in a first and a foreign language: a study of reading comprehension in Japanese (the L1) and English (the L2)*. Lancaster University, University of Lancaster.

Zenisky, A., Hambleton, R. J., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer.

Zhang, X. (2013). The I Don't Know Option in the Vocabulary Size Test. *TESOL Quarterly*, *47*(4), 790–811.

Zhang, X., & Lu, X. (2013). A Longitudinal Study of Receptive Vocabulary Breadth Knowledge Growth and Vocabulary Fluency Development. *Applied Linguistics*, 1–23. http://doi.org/10.1093/applin/amt014

11. Appendices

11.1. Appendix A – Materials for item format pilot study (Chapter 3)

Test A

Choose the right word to go with each meaning. Write the number of that word next to its meaning.

Example item:

1. concrete
2. era
3. fiber
4. hip
5. loop
6. summit

_5_ circular shape

_6_ top of a mountain

_2_ a long period of time

1. alimony
2. beagle
3. counterclaim
4. kestrel
5. proclivity
6. reprise

___ small dog with long ears

___ statement made opposing a previous statement

___ money for the care of children, paid regularly after a divorce

1. cerise
2. jocular
3. lascivious
4. palatial
5. spangled
6. workaday

___ plain and practical

___ bright red in colour

___ covered with small bright decorations

1. aperitif
2. carafe
3. feint
4. gyroscope
5. planetarium
6. riddance

___ drink taken before a meal

___ pretend attack to trick the enemy

___ place where a machine shows the way stars move

Test A

Choose the best meaning for each word. If you do not know the word at all, do not guess. Wrong guesses will be taken away from your correct answers. However, if you think you might know the meaning or part of it, then you should try to find an answer.

Example item:

miniature: It is a miniature.

a a very small thing of its kind

b an instrument for looking at very small objects

c a very small living creature

d a small line to join letters in handwriting

cyberpunk: I like **cyberpunk**.

a medicine that does not use drugs

b one variety of science fiction

c the science of eating

d a society ruled by technical experts

nymphomaniac: She is a **nymphomaniac**.

a person expressing uncontrolled sexual desire

b antisocial person

c innocent rural person

d person who repeats the same crime after punishment

serviette: Where is my **serviette**?

a girl who helps in the house

b piece of glass which makes things look bigger

c large flat plate

d piece of cloth or paper for wiping your mouth

bylaw: They made a **bylaw**.

a publisher's list of older books

b additional rule

c code made of lines, read by machines

d policy that morally condemns people

dachshund: She loves her **dachshund**.

a warm fur hat

b thick floor rug with special patterns

c small dog with short legs and a long back

d old musical instrument with twelve strings

muff: This **muff** belonged to my grandmother.

a tube of animal hair for keeping the hands warm

b cover for a teapot

c long rope of feathers to wear around the neck

d bed cover made from squares of material sewn together

magnanimity: We will never forget her **magnanimity**.

a friendliness

b courage

c generosity

d sincerity

exactitude: She was well known for her **exactitude**.

a courage under pressure

b sense of fairness

c habit of making unreasonable demands

d ability to be very accurate

skylark: We watched a **skylark**.

a show with planes flying in patterns

b human-made object going round the earth

c person who does funny tricks

d small bird that flies high as it sings

Test A

Complete each gap below with the word that best fits the definition. The first letter of the word is given. Make sure to write one letter on each of the spaces.

Example item:

a very small thing of its kind

m *i n i a t u r e*

a furry animal with a long striped tail that looks like a monkey and lives in Madagascar

l __ __ __ __

good smelling substance that comes out of trees

f __ __ __ __ __ __ __ __ __ __ __

low basin for washing the body after using the toilet

b __ __ __ __

large room for eating

r __ __ __ __ __ __ __ __

an animal with a pocket for babies

m __ __ __ __ __ __ __ __

ending of a story which solves the mystery

d __ __ __ __ __ __ __ __ __

weak and soft

e __ __ __ __ __

associated with forests and trees

s __ __ __ __ __

oriented to time and location

s __ __ __ __ __ __ __ __ __ __ __ __ __

Test A

Complete each gap below with the word that best fits the definition. The first letter of the word is given. Make sure to write one letter on each of the spaces.

Example item:

a very small thing of its kind

It is a **m *i n i a t u r e.***

very cheerful and friendly

He was very **j __ __ __ __ __.**

trying to teach people something

Her approach is **d __ __ __ __ __ __ __.**

determined to do something in one's own way

He was a **h __ __ __ __ __ __ __ __ __** child.

crushed together

I **s __ __ __ __ __ __ __ __** the paper up.

stroking and kissing one another

Do you see those couples **c __ __ __ __ __ __ __ __ __** in public?

take care to avoid confrontation

Let's not **p __ __ __ __ __ __ __ __** around.

a repeated short high musical sound

He practised the **t __ __ __ __.**

a plant with large pink, purple, white or yellow flowers growing in groups

This **a __ __ __ __ __** is very pretty.

a desk made to hold a book at a good height for reading to an audience

He stood behind the **l __ __ __ __ __ __.**

Test B

Choose the right word to go with each meaning. Write the number of that word next to its meaning.

Example item:

1. concrete
2. era
3. fiber
4. hip
5. loop
6. summit

_5_ circular shape

_6_ top of a mountain

_2_ a long period of time

1. audacious
2. didactic
3. jovial
4. headstrong
5. morose
6. vindictive

___ cheerful and friendly

___ trying to teach something

___ determined to do something in one's own way

1. azalea
2. fruition
3. lectern
4. spleen
5. trill
6. vestibule

___ repeated high musical sound

___ small plant with many flowers growing in groups

___ desk made to hold a book at a good height for reading

1. berate
2. canoodle
3. lacerate
4. pussyfoot
5. revile
6. scrunch

___ crush together

___ stroke and kiss one another

___ take care to avoid confrontation

Test B

Choose the best meaning for each word. If you do not know the word at all, do not guess. Wrong guesses will be taken away from your correct answers. However, if you think you might know the meaning or part of it, then you should try to find an answer.

Example item:

miniature: It is a miniature.

a a very small thing of its kind

b an instrument for looking at very small objects

c a very small living creature

d a small line to join letters in handwriting

beagle: He owns two **beagles**.

a fast cars with roofs that fold down

b large guns that can shoot many people quickly

c small dogs with long ears

d houses built at holiday places

counterclaim: They made a **counterclaim**.

a a statement opposing a previous statement

b a request for a shop to take back things with faults

c an agreement between two companies to exchange work

d a promise to do something

alimony: The article was about **alimony**.

a feelings of bitterness and annoyance, expressed sharply

b money for the care of children, paid regularly after a divorce

c giving praise for excellent ideas

d a metal which breaks easily and is bluish white

aperitif: He had an **aperitif**.

a a long chair for lying on

b a private singing teacher

c a large hat with tall feathers

d a drink taken before a meal

feint: He made a **feint**.

a small cake with dried fruit

b thing with wheels for moving heavy objects

c pretend attack or move to trick the enemy

d serious mistake

planetarium: The **planetarium** was interesting.

a place where planes are built

b place where a machine shows the way stars move

c course to teach people good planning skills

d place where fish are kept

workaday: These are **workaday** clothes.

a plain and practical

b suitable for parties after work

c old and worn out

d made to be thrown away after each working day

cerise: Her skirt was **cerise**.

a a bright red colour

b made of a thin, soft material

c a pale blue-green colour

d made of expensive fabric with pretty patterns

spangled: Her dress was **spangled**.

a torn into thin strips

b covered with small bright decorations

c made with lots of folds of fabric

d ruined by touching something very hot

Test C

Complete each gap below with the word that best fits the definition. The first letter of the word is given. Make sure to write one letter on each of the spaces.

Example item:

a very small thing of its kind

m *i* *n* *i* *a* *t* *u* *r* *e*

a variety of science fiction where stories are set in a world controlled by technology

c __ __ __ __ __ __ __ __

a person expressing uncontrolled sexual desire

n __ __ __ __ __ __ __ __ __ __ __

piece of cloth or paper for wiping your mouth

s __ __ __ __ __ __ __ __

additional rule made by a local authority applying only to that area

b __ __ __ __

a small dog with short legs and a long back

d __ __ __ __ __ __ __ __

tube of animal hair for keeping the hands warm

m __ __ __

generosity

m __ __ __ __ __ __ __ __ __ __

ability to be very accurate

e __ __ __ __ __ __ __ __ __

small bird that flies high as it sings

s __ __ __ __ __ __

Test C

Complete each gap below with the word that best fits the definition. The first letter of the word is given. Make sure to write one letter on each of the spaces.

Example item:

a very small thing of its kind

It is a **miniature.**

a furry animal with a long striped tail that looks like a monkey and lives in Madagascar

We saw a **l** __ __ __ __.

good smelling substance that comes out of trees

He brought some **f** __ __ __ __ __ __ __ __ __ __ __.

low basin for washing the body after using the toilet

They have a **b** __ __ __ __.

large room for eating

We met in the **r** __ __ __ __ __ __ __ __ of the school.

an animal with a pocket for babies

A **m** __ __ __ __ __ __ __ __ lives in Australia.

ending of a story which solves the mystery

I was disappointed with the **d** __ __ __ __ __ __ __ __ __.

weak and soft

He has become **e** __ __ __ __ __.

associated with forests and trees

The painting had a **s** __ __ __ __ __ theme.

oriented to time and location

My theory is **s** __ __ __ __ __ __ __ __ __ __ __ __ __.

Test C

Choose the right word to go with each meaning. Write the number of that word next to its meaning.

Example item:

1. concrete
2. era
3. fiber
4. hip
5. loop
6. summit

_5_ circular shape

_6_ top of a mountain

_2_ a long period of time

1. balaclava
2. bidet
3. frankincense
4. lemur
5. sirloin
6. trilby

___ furry animal with a long tail

___ good smelling substance that comes out of trees

___ low basin for washing the body after using the toilet

1. apoplectic
2. effete
3. limpid
4. spatiotemporal
5. sylvan
6. unctuous

___ weak and soft

___ associated with forests

___ oriented to time and location

1. denouement
2. epidermis
3. fedora
4. marsupial
5. novella
6. refectory

___ room for eating

___ an animal with a pocket for babies

___ ending of a story which solves the mystery

Test C

Choose the best meaning for each word. If you do not know the word at all, do not guess. Wrong guesses will be taken away from your correct answers. However, if you think you might know the meaning or part of it, then you should try to find an answer.

Example item:

miniature: It is a miniature.

a a very small thing of its kind

b an instrument for looking at very small objects

c a very small living creature

d a small line to join letters in handwriting

jovial: He was very **jovial**.

a low on the social scale

b likely to criticize others

c cheerful and friendly

d interesting or exciting

didactic: Her approach is **didactic**.

a trying to teach something

b difficult to believe

c about exciting actions

d unclear in meaning

headstrong: He was a **headstrong** child.

a very clever

b given too many good things

c difficult to keep quiet

d determined to do what it wants

scrunch: It was **scrunched** up.

a done with many mistakes

b crushed together

c cut into rough pieces

d thrown violently into the air

canoodle: Do you see that couple **canoodling**?

a spreading false and evil ideas about others

b looking for a free meal

c merging into the crowd

d stroking and kissing one another

pussyfoot: Let's not **pussyfoot** around.

a criticise unreasonably

b take care to avoid confrontation

c attack indirectly

d suddenly start

trill: He practised the **trill**.

a type of stringed instrument

b repeated high musical sound

c way of throwing the ball

d dance step of turning round very fast on the toes

azalea: This **azalea** is very pretty.

a sea shell shaped like a fan

b light natural fabric

c long piece of material worn in India

d small plant with many flowers growing in groups

lectern: He stood behind the **lectern**.

a desk made to hold a book at a good height for reading

b table or block used for church ceremonies

c place where you buy drinks

d heavy door made of wood

Test C

Complete each gap below with the word that best fits the definition. The first letter of the word is given. Make sure to write one letter on each of the spaces.

Example item:

a very small thing of its kind

m *i n i a t u r e*

a small dog with short legs and long ears, used in hunting

b __ __ __ __ __

a statement made opposing a previous statement

c __ __ __ __ __ __ __ __ __ __ __

money for the care of children, paid regularly after a divorce

a __ __ __ __ __ __

drink taken before a meal

a __ __ __ __ __ __ __

pretend attack or move to trick the enemy

f __ __ __ __

place where a machine shows the way stars move

p __ __ __ __ __ __ __ __ __ __

plain and practical, ordinary

w __ __ __ __ __ __ __

bright red in colour

c __ __ __ __ __

covered with small bright decorations

s __ __ __ __ __ __ __

Test C

Complete each gap below with the word that best fits the definition. The first letter of the word is given. Make sure to write one letter on each of the spaces.

Example item:

a very small thing of its kind

It is a m *i n i a t u r e.*

a variety of science fiction where stories are set in a world controlled by technology

I like **c** __ __ __ __ __ __ __ __ novels.

a person expressing uncontrolled sexual desire

She is a **n** __ __ __ __ __ __ __ __ __ __ __.

piece of cloth or paper for wiping your mouth

Where is my **s** __ __ __ __ __ __ __ __?

additional rule made by a local authority applying only to that area

They made a **b** __ __ __ __.

a small dog with short legs and a long back

She loves her **d** __ __ __ __ __ __ __ __.

tube of animal hair for keeping the hands warm

This **m** __ __ __ belonged to my grandmother.

generosity

We will never forget her **m** __ __ __ __ __ __ __ __ __ __.

ability to be very accurate

She was well known for her **e** __ __ __ __ __ __ __ __ __.

small bird that flies high as it sings

We watched a **s** __ __ __ __ __ __.

Test D

Choose the right word to go with each meaning. Write the number of that word next to its meaning.

Example item:

1. concrete
2. era
3. fiber
4. hip
5. loop
6. summit

5 circular shape

6 top of a mountain

2 a long period of time

1. cyberpunk
2. nymphomaniac
3. serviette
4. superscript
5. tipster
6. wigwam

___ a variety of science fiction

___ a person expressing uncontrolled sexual desire

___ piece of cloth or paper for wiping your mouth

1. demeanour
2. exactitude
3. magnanimity
4. scrimmage
5. skylark
6. tamarisk

___ generosity

___ ability to be very accurate

___ small bird that flies high as it sings

1. bylaw
2. dachshund
3. furlough
4. gecko
5. muff
6. zeitgeist

___ additional rule

___ small dog with short legs and a long back

___ tube of animal hair for keeping the hands warm

Test D

Choose the best meaning for each word. If you do not know the word at all, do not guess. Wrong guesses will be taken away from your correct answers. However, if you think you might know the meaning or part of it, then you should try to find an answer.

Example item:

miniature: It is a miniature.

a a very small thing of its kind

b an instrument for looking at very small objects

c a very small living creature

d a small line to join letters in handwriting

lemur: We saw a **lemur**.

a priest from an eastern religion

b person with a very bad skin disease

c furry animal with a long tail

d purple fish from hot countries

frankincense: He brought some **frankincense**.

a sweet smelling white flowers

b soft cheese made in France

c food made from yellow coloured rice and shellfish

d good smelling substance that comes out of trees

bidet: They have a **bidet**.

a low basin for washing the body after using the toilet

b large fierce brown dog

c small private swimming pool

d man to help in the house

refectory: We met in the **refectory**.

a room where legal papers can be signed

b room for eating

c room for several people to sleep in

d room with glass walls

marsupial: It is a **marsupial**.

a an animal with hard feet

b a plant that takes several years to grow

c a plant with flowers that turn to face the sun

d an animal with a pocket for babies

denouement: I was disappointed with the **denouement**.

a small place to live which is part of a bigger building

b amount of money paid for a piece of work

c ending of a story which solves the mystery

d official report of the results of a political meeting

effete: He has become **effete**.

a weak and soft

b too fond of strong drink

c unable to leave his bed

d extremely easy to annoy

sylvan: The painting had a **sylvan** theme.

a lost love

b wandering

c forest

d casual folk

spatiotemporal: My theory is **spatiotemporal**.

a focused on small details

b annoying to people

c objectionably modern

d oriented to time and location

Test D

Complete each gap below with the word that best fits the definition. The first letter of the word is given. Make sure to write one letter on each of the spaces.

Example item:

a very small thing of its kind

m *i n i a t u r e*

very cheerful and friendly

j __ __ __ __ __

trying to teach people something

d __ __ __ __ __ __ __

determined to do something in one’s own way

h __ __ __ __ __ __ __ __ __

crushed together

s __ __ __ __ __ __ __ __

stroking and kissing one another

c __ __ __ __ __ __ __ __ __

take care to avoid confrontation

p __ __ __ __ __ __ __ __

a repeated short high musical sound

t __ __ __ __

a plant with large pink, purple, white or yellow flowers growing in groups

a __ __ __ __ __

a desk made to hold a book at a good height for reading to an audience

l __ __ __ __ __ __

Test D

Complete each gap below with the word that best fits the definition. The first letter of the word is given. Make sure to write one letter on each of the spaces.

Example item:

a very small thing of its kind

It is a **m** *i n i a t u r e.*

a small dog with short legs and long ears, used in hunting

He owns a **b** __ __ __ __ __.

a statement opposing a previous statement

They made a **c** __ __ __ __ __ __ __ __ __ __ __.

money for the care of children, paid regularly after a divorce

The article was about **a** __ __ __ __ __ __.

drink taken before a meal

He had an **a** __ __ __ __ __ __ __.

pretend attack or move to trick the enemy

He made a **f** __ __ __ __.

place where a machine shows the way stars move

The **p** __ __ __ __ __ __ __ __ __ __ was interesting.

plain and practical, ordinary

These are **w** __ __ __ __ __ __ __ clothes.

bright red in colour

Her skirt was **c** __ __ __ __ __.

covered with small bright decorations

Her dress was **s** __ __ __ __ __ __ __.

TARGET	Meaning (minimally required recall in 1st line, additional optional information in 2nd line)
didactic	trying to teach something
jovial	very cheerful and friendly
headstrong	determined to do something in one’s own way refusing to listen to advice
trill	a repeated short high (musical) sound
azalea	a plant or bush with large (pink, purple, white or yellow) flowers growing in groups grown in a pot or in a garden
lectern	a desk/stand made to hold a book/notes at a good height for reading to an audience (in church, giving a talk, etc.)
beagle	a small dog with short legs and long ears used in hunting
alimony	money for the care of children, paid (regularly) after a divorce
counterclaim	a statement/demand made opposing a previous statement
denouement	ending of a story/play/book which solves/explains/settles everything/the mystery the end result of a situation; part of the story after the climax
marsupial	an animal with a pocket for babies
refectory	large room for eating usually in religious institutions or schools
scrunched	crushed together to squeeze something into a small round shape in your hands, to make something become smaller
pussyfoot	take care to avoid confrontation/upsetting anyone
canoodling	stroking and kissing one another
spangled	covered/decorated with small bright decorations/shiny things
cerise	bright red/pinkish-red in colour
workaday	ordinary, plain practical
sylvan	associated with forests or trees
effete	weak/soft, without the power that it once had OR looking or behaving like a woman
spatiotemporal	oriented to time and location
aperitif	drink taken before a meal usually contains alcohol
feint	pretend attack or move to trick the enemy used in fights/wars/sports
planetarium	place that shows the way stars move place where a machine shows the movements in the universe
muff	tube for putting your hands into to keep them warm made of animal hair or other warm material
bylaw	additional rule/regulation made by a local authority applying only to that area
dachshund	a small dog with short legs and a long back long ears
magnanimity	Kindness OR generosity OR forgivingness
skylark	a small bird that sings while it flies high up in the sky
exactitude	ability to be very accurate
bidet	low basin for washing the body used after using the toilet
frankincense	good smelling substance that comes out of trees OR substance that is burnt to give a pleasant smell especially used during religious ceremonies
lemur	a furry animal that looks like a monkey + with a long striped tail OR lives in Madagascar
serviette	piece of cloth or paper for wiping your mouth
cyberpunk	a variety of science fiction where stories are set in a world controlled by technology
nymphomaniac	a person (female) expressing uncontrolled sexual desire

11.2. Appendix B – Materials for item format main study (Chapter 3)

11.2.1. Items for test of receptive derivative knowledge

[N.B. The solution (baseword according to Nation's BNC lists) provided in column 1 was not visible to candidates.]

Write down the word which you think is the basis of the words displayed. Do not just copy one of the three words.

Example:
regretfully
regrettable
regretting
Answer:
regret

accurate	inaccuracy	accurately	accuracies
behaviour	misbehaviour	behavioural	behaviorist
blend	blender	blending	unblended
collaborate	collaboration	collaboratively	collaborator
controversy	uncontroversial	controversially	controversies
document	documentation	undocumented	documenting
draft	redrafted	drafting	drafter
encounter	encounters	encountered	encountering
exception	exceptionable	exceptionally	exceptionalities
fertile	infertility	fertiliser	fertilizing
glow	glowingly	glowed	glowing
grateful	gratefulness	ungrateful	gratefully
immune	immunity	immunized	immunising
initiate	initiative	initiation	uninitiated
justify	justifiably	justification	unjustified
margin	marginal	marginalized	marginally
mortal	mortalilty	immortal	immortally
motive	motivation	motiveless	motivating
negotiate	negotiation	negotiator	renegotiating
oblige	nonobligatory	obligingly	obligation
palm	palmed	palming	palms
pepper	peppery	peppers	peppered
phrase	phrasal	rephrase	misphrasing
predict	predictability	prediction	unpredictably
preserve	preservation	preservative	preserving
quantity	quantitative	quantities	quantitatively
remedy	remedial	remedied	remedies
resemble	resemblance	resembling	resembles
ritual	ritually	ritualistic	ritualisation
structure	restructuring	poststructuralism	structurally
suburb	suburban	suburbs	suburbanisation
summary	summarise	summarisation	summarily
universe	universal	universally	universalisation
vulnerable	invulnerability	vulnerably	invulnerable
withdraw	withdrawal	withdrawing	withdrew
youth	youthful	youthfully	youthfulness

11.2.2. Items for test of form-meaning link knowledge

MC items

1. behaviour I don't like his **behaviour**.
    a) way of doing things
    b) personal possessions
    c) attitude towards women
    d) sense of fashion

2. pepper I would like some **pepper**.
    a) substance to clean clothes
    b) powder to make food hot
    c) material used for writing on
    d) information about a course

3. summary He gave me a **summary**.
    a) long story about warmest season of the year
    b) small telephone that can be carried around
    c) short description that gives the main facts
    d) instrument to look at small objects

4. document Is this the right **document**?
    a) answer to a difficult question
    b) space to park your car
    c) person with a medical degree
    d) official paper

5. remedy She took a **remedy**.
    a) medicine to cure a disease
    b) train leaving late at night
    c) picture to remind her of a special event
    d) chance to appear in the media

6. youth He enjoyed his **youth**.
    a) drink served before a meal
    b) time of life when he was young
    c) game involving two teams and a small ball
    d) expensive sweet fruit

7. quantity I liked the **quantity** of food.
    a) standard
    b) taste
    c) amount
    d) smell

8. structure I can't see any **structure**.
    a) area in a large public place where people can meet
    b) point where something changes
    c) art object made of wood
    d) way in which the parts of something are organised

9. universe — Let us look at the **universe**.

a) system of stars and planets
b) institution at the highest level of education
c) organization of workers
d) particular group of people all wearing the same clothes

10. controversy — This will cause a big **controversy**.

a) physical fight between many people
b) accident or explosion
c) problem that you use mathematics to solve
d) public disagreement

11. motive — I like his **motive**.

a) formal proposal
b) large car for carrying goods
c) reason for doing something
d) attitude towards politics

12. ritual — This is our **ritual**.

a) deep dish for food
b) person who helps with cleaning
c) something that is done regularly
d) piece of kitchen equipment to keep things cold

13. draft — This is only a **draft**.

a) copy of the original
b) rough unfinished version
c) little stain of dirt
d) small amount of money

14. palm — I showed him my **palm**.

a) collection of soft toy animals
b) machine that makes things look bigger
c) drawing of an island
d) inner surface of the hand

15. suburb — She liked the **suburb**.

a) ship that can travel underwater
b) *area on the edge of a large town*
c) type of alcoholic drink
d) time when she was not working

16. exception — This is an **exception**.

a) device for giving light
b) thing that does not follow a rule
c) sad expression on someone's face
d) general agreement that something is right

17. margin There is a big **margin**.

a) celebration of a couple's relationship
b) event in which many people walk through a public place
c) plate of food containing a lot of fat
d) space at the side of a printed page

18. phrase I like this **phrase**.

a) movie about characters with special strengths
b) group of words which has a particular meaning
c) first stage in a series of events
d) competition in which everyone tries to be the fastest

19. collaborate I like to **collaborate** with him.

a) play
b) talk
c) work
d) sing

20. glow Can you see it **glow**?

a) almost fall
b) move into a low position
c) produce a soft light
d) be less active

21. predict He thinks he can **predict** events.

a) say what will happen
b) organise on his own
c) change what will happen
d) cancel on his own

22. encounter When did you **encounter** him?

a) meet
b) tell
c) attack
d) answer

23. negotiate We had to **negotiate**.

a) make it clear to everyone
b) say no to a suggestion
c) claim it without proof
d) try to reach an agreement

24. preserve I try to **preserve** it.

a) introduce
b) keep it in its original state
c) help achieve
d) discover by a science experiment

25. initiate I will **initiate** it.

a) steal
b) start
c) repair
d) describe

26. justify Can you **justify** this?

a) show that it is right
b) argue against the facts that support it
c) call it out loudly
d) make it legal

27. resemble He says that we **resemble** each other.

a) look like
b) meet again
c) really love
d) have the right character for

28. blend I will **blend** them.

a) make unable to see
b) throw away
c) mix together
d) tell in great detail

29. oblige She wanted to **oblige** her to do it.

a) force
b) pay
c) ask
d) allow

30. withdraw I want to **withdraw**!

a) make a picture of something
b) continue doing it
c) move back or away from a situation
d) admit that I have lost the competition

31. accurate This is **accurate**.

a) very noisy
b) correct in every detail
c) old and broken
d) good for your health

32. immune He was **immune** to it.

a) showing no interest
b) completely uncertain
c) protected from it and therefore able to avoid it
d) not having much experience

33. mortal He is **mortal**.

a) aggressive and violent
b) easily embarrassed
c) behaving in a correct and honest way
d) unable to continue living forever

34. fertile   It looks very **fertile**.
- a) able to produce a lot of healthy plants
- b) more expensive than necessary
- c) enjoyable and attractive
- d) frightening and violent

35. grateful   They were **grateful**.
- a) helpful
- b) careful
- c) beautiful
- d) thankful

36. vulnerable   He was **vulnerable**.
- a) very interested
- b) convinced of his abilities
- c) weak and easily hurt
- d) famous

MM items

1. acquisition
2. behaviour   ___ way of doing things
3. layer   ___ powder to make food hot
4. pepper   ___ short description that gives the main facts
5. summary
6. tube

1. conflict
2. document   ___ official paper
3. fate   ___ medicine to cure a disease
4. passenger   ___ time of life when a person is young
5. remedy
6. youth

1. curtain
2. highway   ___ amount of something
3. liberty   ___ system of stars and planets
4. quantity   ___ way in which the parts of something are organised
5. structure
6. universe

1. controversy
2. motive   ___ public disagreement
3. ritual   ___ reason for doing something
4. solution   ___ something that is done regularly
5. tragedy
6. wealth

1. bench	
2. cattle	___ rough unfinished version
3. draft	___ inner surface of the hand
4. era	___ area on the edge of a large town
5. palm	
6. suburb	

1. exception	
2. infant	___ thing that does not follow a rule
3. margin	___ space at the side of a printed page
4. notion	___ group of words which has a particular meaning
5. phrase	
6. prospect	

1. collaborate	
2. glow	___ produce a soft light
3. imply	___ say what will happen
4. launch	___ work together with somebody
5. offend	
6. predict	

1. encounter	
2. interfere	___ try to reach an agreement
3. merge	___ keep something in its original state
4. negotiate	___ meet somebody or discover something
5. preserve	
6. render	

1. collapse	
2. explore	___ make something begin
3. initiate	___ show that something is right
4. justify	___ look like another person or thing
5. resemble	
6. succeed	

1. blend	
2. lease	___ mix together
3. manufacture	___ force somebody to do something
4. oblige	___ move back or away from a situation
5. reject	
6. withdraw	

1. accurate	
2. immune	___ correct in every detail
3. mortal	___ unable to continue living forever
4. mutual	___ protected from something and therefore able to avoid it
5. unique	
6. voluntary	

1. ancient
2. fertile
3. grateful
4. profound
5. supreme
6. vulnerable

___ weak and easily hurt

___ feeling or showing thanks

___ able to produce a lot of healthy plants

RECALL ITEMS (DEFINITION)

the way someone does or says things

b _ _ _ _ _ _ _ _

a short description that gives the main facts

s _ _ _ _ _ _

a powder made from dried seeds to make food hot

p _ _ _ _ _

an area on the edge of a large city where people who work in the city often live

s _ _ _ _ _

a rough version of something that is not yet in its final form

d _ _ _ _

the inner surface of the hand

p _ _ _

an official paper

d _ _ _ _ _ _ _

a treatment or medicine to cure a disease

r _ _ _ _ _

the time of life when a person is young

y _ _ _ _

a thing that does not follow a rule

e _ _ _ _ _ _ _ _

a group of words which have a particular meaning when used together

p _ _ _ _ _

the empty space at the side of a written or printed page

m _ _ _ _ _

an amount of something

q _ _ _ _ _ _ _

the system of stars and planets in space

u _ _ _ _ _ _ _

the way in which the parts of something are organised

s _ _ _ _ _ _ _ _

to work together with somebody

c _ _ _ _ _ _ _ _ _ _

to say that something will happen in the future

p _ _ _ _ _ _

to produce a soft, warm light

g _ _ _

a reason for doing something

m _ _ _ _ _

public discussion about something that people strongly disagree about

c _ _ _ _ _ _ _ _ _ _

something that is done regularly and always in the same way

r _ _ _ _ _

to meet somebody, or discover or experience something

e _ _ _ _ _ _ _ _

to try to reach an agreement by discussion

n _ _ _ _ _ _ _ _

to keep something in its original state

p _ _ _ _ _ _ _

to make something begin

i _ _ _ _ _ _ _

to show that somebody/something is right

j _ __ _ _ _

to look like or be similar to another person or thing

r _ _ _ _ _ _ _

correct and true in every detail

a _ _ _ _ _ _ _

protected from something and therefore able to avoid it

i _ _ _ _ _

unable to continue living forever

m _ _ _ _ _

to move back or away from a situation

w _ _ _ _ _ _ _

to mix two or more things together

b _ _ _ _

to force somebody to do something

o _ _ _ _ _

able to produce a lot of healthy plants

f _ _ _ _ _ _

feeling or showing thanks

g _ _ _ _ _ _ _

weak and easily hurt physically or emotionally

v _ _ _ _ _ _ _ _ _

RECALL ITEMS (CONTEXT)

the way someone does or says things

I don’t like his b _ _ _ _ _ _ _ _.

a short description that gives the main facts

He gave me a s _ _ _ _ _ _.

a powder made from dried seeds to make food hot

I would like some p _ _ _ _ _.

a reason for doing something

I like his m _ _ _ _ _.

public discussion about something that people strongly disagree about

This will cause a big c _ _ _ _ _ _ _ _ _ _ .

something that is done regularly and always in the same way

This is our r _ _ _ _ _ .

an official paper

Is this the right d _ _ _ _ _ _ _.

a treatment or medicine to cure a disease

She took a r _ _ _ _ _.

the time of life when a person is young

He enjoyed his y _ _ _ _

an area on the edge of a large city where people who work in the city often live

She liked the s _ _ _ _ _.

a rough version of something that is not yet in its final form

This is only a d _ _ _ _.

the inner surface of the hand

I showed him my p _ _ _.

an amount of something

I liked the q _ _ _ _ _ _ _ of food.

the system of stars and planets in space

Let us look at the u _ _ _ _ _ _ _.

the way in which the parts of something are organised

I can’t see any s _ _ _ _ _ _ _ _ .

a thing that does not follow a rule

This is an e _ _ _ _ _ _ _ _.

a group of words which have a particular meaning when used together

I like this p _ _ _ _ _.

the empty space at the side of a written or printed page

There is a big m _ _ _ _ _.

to work together with somebody

I like to c _ _ _ _ _ _ _ _ _ _ with him.

to say that something will happen in the future

He thinks he can p _ _ _ _ _ _ events.

to produce a soft, warm light

Can you see it g _ _ _?

to meet somebody, or discover or experience something

When did you e _ _ _ _ _ _ _ _ him.

to try to reach an agreement by discussion

We had to n _ _ _ _ _ _ _ _.

to keep something in its original state

I try to p _ _ _ _ _ _ _ it.

to make something begin

I will i _ _ _ _ _ _ _ it.

to show that somebody/something is right

Can you j _ _ _ _ _ _ this?

to look like or be similar to another person or thing

He said that we r _ _ _ _ _ _ _ each other.

to move back or away from a situation

I want to w _ _ _ _ _ _ _!

to mix two or more things together

I will b _ _ _ _ them.

to force somebody to do something

She wanted to o _ _ _ _ _ her to do it.

correct and true in every detail

This is a _ _ _ _ _ _ _.

protected from something and therefore able to avoid it

He was i _ _ _ _ _ to it.

unable to continue living forever

He is m _ _ _ _ _.

able to produce a lot of healthy plants

It looks very f _ _ _ _ _ _ .

feeling or showing thanks

They were g _ _ _ _ _ _ _.

weak and easily hurt physically or emotionally

He was v _ _ _ _ _ _ _ _ _.

11.2.3. Items for test of collocation knowledge

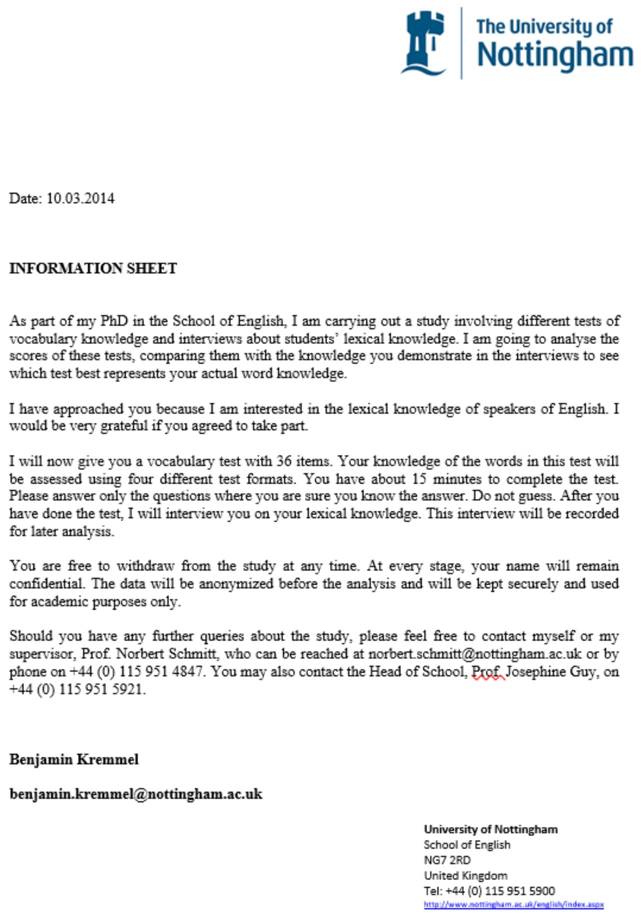
Out of the three options, choose the <u>two most natural and frequent word combinations</u>. Choose ONLY TWO of the options. You must tick BOTH correct combinations to get the point.

Example:

☐ make homework ☒ do homework ☒ complete homework

aggressive behaviour	antisocial behaviour	ugly behaviour
boiled pepper	ground pepper	crushed pepper
brief summary	quick summary	small summary
unpublished document	broad document	signed document
useful remedy	traditional remedy	popular remedy
misspent youth	nasty youth	homeless youth
huge quantity	little quantity	unknown quantity
basic structure	organisational structure	faulty structure
mighty universe	expanding universe	entire universe
stir controversy	provoke controversy	produce controversy
primary motive	casual motive	possible motive
regular ritual	religious ritual	nightly ritual
ideal draft	revised draft	final draft
sweaty palm	closed palm	cupped palm
middle-class suburb	wealthy suburb	good suburb
notable exception	possible exception	portable exception
slim margin	short margin	narrow margin
coin a phrase	borrow a phrase	say a phrase
collaborate with	collaborate on	collaborate at
shiny glow	golden glow	faint glow
predict the outcome	predict the weather	predict the environment
encounter disasters	encounter difficulties	encounter problems
negotiate prices	negotiate discussions	negotiate contracts
preserve nature	preserve love	preserve peace
initiate a business	initiate a conversation	initiate a process
justify actions	justify claims	justify morals
resemble closely	resemble exactly	resemble strongly
blend colours	blend looks	blend ingredients
happy to oblige	pleased to oblige	willing to oblige
withdraw troops	withdraw money	withdraw pictures
accurate punishment	accurate measurement	accurate description
fully immune	relatively immune	largely immune
mortal fight	mortal sin	mortal enemy
fertile soil	fertile floor	fertile ground
deeply grateful	highly grateful	extremely grateful
particularly vulnerable	especially vulnerable	greatly vulnerable

11.3. Appendix C – Consent forms for item format pilot study (Chapter 3)



The University of Nottingham

Date: 10.03.2014

INFORMATION SHEET

As part of my PhD in the School of English, I am carrying out a study involving different tests of vocabulary knowledge and interviews about students' lexical knowledge. I am going to analyse the scores of these tests, comparing them with the knowledge you demonstrate in the interviews to see which test best represents your actual word knowledge.

I have approached you because I am interested in the lexical knowledge of speakers of English. I would be very grateful if you agreed to take part.

I will now give you a vocabulary test with 36 items. Your knowledge of the words in this test will be assessed using four different test formats. You have about 15 minutes to complete the test. Please answer only the questions where you are sure you know the answer. Do not guess. After you have done the test, I will interview you on your lexical knowledge. This interview will be recorded for later analysis.

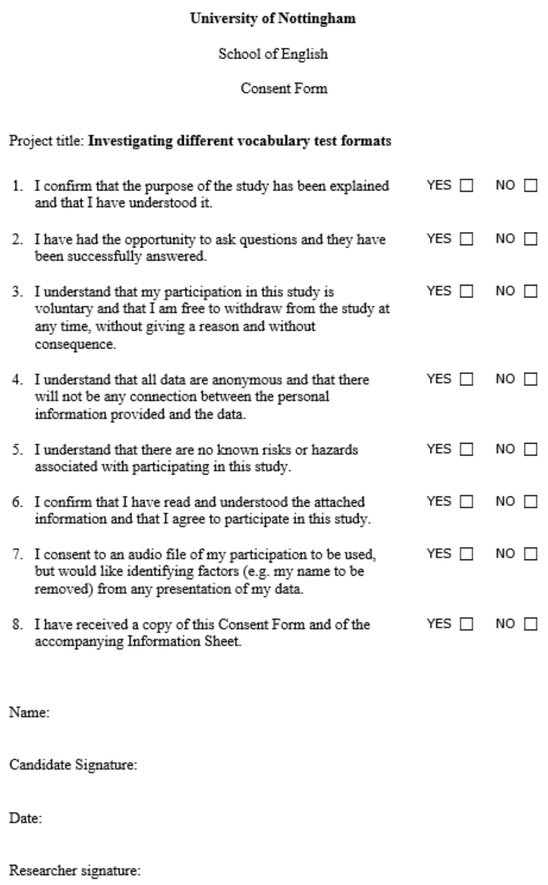
You are free to withdraw from the study at any time. At every stage, your name will remain confidential. The data will be anonymized before the analysis and will be kept securely and used for academic purposes only.

Should you have any further queries about the study, please feel free to contact myself or my supervisor, Prof. Norbert Schmitt, who can be reached at norbert.schmitt@nottingham.ac.uk or by phone on +44 (0) 115 951 4847. You may also contact the Head of School, Prof. Josephine Guy, on +44 (0) 115 951 5921.

Benjamin Kremmel

benjamin.kremmel@nottingham.ac.uk

University of Nottingham
School of English
NG7 2RD
United Kingdom
Tel: +44 (0) 115 951 5900
http://www.nottingham.ac.uk/english/index.aspx



University of Nottingham

School of English

Consent Form

Project title: **Investigating different vocabulary test formats**

1. I confirm that the purpose of the study has been explained and that I have understood it. YES □ NO □

2. I have had the opportunity to ask questions and they have been successfully answered. YES □ NO □

3. I understand that my participation in this study is voluntary and that I am free to withdraw from the study at any time, without giving a reason and without consequence. YES □ NO □

4. I understand that all data are anonymous and that there will not be any connection between the personal information provided and the data. YES □ NO □

5. I understand that there are no known risks or hazards associated with participating in this study. YES □ NO □

6. I confirm that I have read and understood the attached information and that I agree to participate in this study. YES □ NO □

7. I consent to an audio file of my participation to be used, but would like identifying factors (e.g. my name to be removed) from any presentation of my data. YES □ NO □

8. I have received a copy of this Consent Form and of the accompanying Information Sheet. YES □ NO □

Name:

Candidate Signature:

Date:

Researcher signature:

11.4. Appendix D – Consent form for item format main study (Chapter 3)

Instruktionen

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Im Zuge meines Doktoratsstudiums an der Universität Nottingham, UK, führe ich eine Studie zur Aussagekraft verschiedener Vokabeltestformate durch. Ich bitte Sie daher den folgenden Onlinetest gewissenhaft und ohne Hilfsmittel zu bearbeiten.

Der Test hat 7 Teile. Sie haben eine Stunde Zeit, diese Tests auszufüllen. Raten Sie dabei nicht, sondern beantworten Sie nur Fragen, bei deren Antwort Sie sich einigermassen sicher sind. Wenn Sie einen Testteil abgeschlossen haben, gehen Sie NICHT zurück, um Ihre vorherigen Antworten zu überprüfen oder zu ändern.

Nach dem Ende der Tests wäre ich Ihnen dankbar wenn Sie einen kurzen Fragebogen zu Ihrer Person ausfüllen könnten. Ihre Daten werden noch vor der Analyse anonymisiert und zu jeder Zeit vertraulich behandelt und für Dritte unzugänglich verwahrt.

Sie können jederzeit von der Teilnahme an dieser Studie zurücktreten. Wenn Sie weitere Fragen zur Studie haben kontaktieren Sie bitte mich unter unten angegebener Mailadresse oder meinen Betreuungsprofessor Norbert Schmitt unter norbert.schmitt@nottingham.ac.uk.

Vielen Dank für Ihre Mithilfe!

Benjamin Kremmel

benjamin.kremmel@nottingham.ac.uk

University of Nottingham
School of English
NG7 2RD
United Kingdom
Tel: +44 (0) 115 951 5900
http://www.nottingham.ac.uk/english/index.aspx

Mit dem Start der Umfrage bestätigen Sie, dass sie

- die obige Information gelesen und den Zweck der Studie verstanden haben.
- verstanden haben, dass diese Form der Studie keinerlei bekannte Risiken birgt.
- verstanden haben, dass die Teilnahme an dieser Studie freiwillig ist und Sie jederzeit ohne Angabe von Gründen und ohne Konsequenzen von Ihrer Teilnahme zurücktreten können.
- verstanden haben, dass alle Daten anonymisiert und vertraulich behandelt werden und keine Verbindung der Daten zu Angaben Ihrer Person herstellbar sein wird.
- die Gelegenheit hatten, Fragen zu stellen und diese zu Ihrer Zufriedenheit beantwortet wurden.
- sich bereit erklären, an dieser Studie teilzunehmen.

11.5. Appendix E – Vocabulary items for population identification study (Chapter 4)

1K

receive: You will receive it.

a see

b get

c feel

d become

society: This is our society.

a fast car with no roof

b room for eating

c large group of people

d dog kept as a pet

action: Think about his action.

a something that he did

b performance in a play

c detailed instruction

d way in which he speaks

break: This will break.

a stop working because it is damaged

b become much smoother

c begin to go more quickly

d prepare for something unpleasant

official: It will be official soon.

a turned into a room for people to work

b cooked perfectly

c old and worn out

d approved by the government

activity: I like this activity.

a someone who plays in films

b large hat with tall feathers

c organized event

d political speech

building: This is a new building.

a house

b picture

c form of education

d way of paying bills

especially: I made it especially for you.

a cheaply

b beautifully

c particularly

d carefully

carry: I can carry it.

a drive from one place to another

b hold with my hands

c make more spicy

d read without mistakes

recent: This was a recent event.

a involving a lot of competition

b expensive and enjoyable

c taking place under fair conditions

d happening a short time ago

2K

democracy: This is a democracy.

a characteristic of the people who live in an area

b system of government where people elect leaders

c example of a product to make people buy it

d political march to protest or change a system

aware: I am aware of this.

a in a different place

b know about it

c having strong feelings

d experiencing it soon

flower: I like this flower.

a powder used to make bread

b bird that is kept for its eggs

c object that moves through air

d coloured part of a plant

key: Can you give me the key?

a piece of metal for locking doors

b piece of wood at the bottom of a boat

c piece of electrical musical equipment

d piece of meat cooked on a thin stick

pair: What a nice pair.

a jewellery made of white, round objects

b oval-shaped, green or yellow fruit

c two things that are similar and go together

d container with an open top and a handle cut: This is a big cut.

a opening made with a sharp tool
b small animal with fur kept as a pet
c bed with high sides for a baby
d collection of things in a container

depend: I depend on it.

a make worse
b look down
c keep in a small area
d need its help

daily: Do you do this daily?

a with milk
b every day
c in an attractive way
d irregularly

demand: There was much demand.

a strong request
b confusion
c illness leading to death
d evil spirit

fully: I can understand it fully.

a incorrectly
b angrily
c completely
d partly

3K

reputation: He has a reputation.

a serious disease
b long gun that fires small, metal balls
c opinion that people have about someone
d relationship with a co-worker

pure: This is pure.

a clean and healthy
b low quality
c open to everyone
d extremely silly

fellow: He talked to his fellow students.

a sharing your interests or situation
b frightened or worried
c best performing
d becoming less in number

crop: There is more crop.

a plant such as a grain, fruit, or vegetable
b solid waste from animals
c curved piece of bread eaten for breakfast
d sea creature with ten legs

ingredient: Please hand me the last ingredient.

a bottle of drink with alcohol
b book about cooking
c one of the different foods that another food is made from
d tool used to show if a surface is smooth

grandmother: She is my grandmother.

a mother of my mother
b sister of my mother
c aunt of my mother
d cousin of my mother

employment: I cannot find employment.

a long chair for lying on
b paid work for a company
c a very large cup
d experienced person who gives help

install: I need to install this.

a make it ready to use
b keep for later
c buy from someone
d stop the progress

literally: I mean this literally.

a relating to the sides of an object
b relating to literature
c relating to large amounts
d relating to its original sense

accompany: Will you accompany me?

a help organize the selling of goods
b tell someone something important
c go somewhere with someone
d help in committing a crime

4K

pill: The pill is on the table.

a soft object that you rest your head on in bed
b large number of objects on top of each other
c something that everyone wants
d small piece of medicine that you swallow

log: This log looks funny.

a thick piece of wood cut from a tree

b container that is opened with a key

c large area of water with land all around it

d bread that has been baked in one large piece

developer: She was the first developer.

a someone who does not eat meat or fish

b someone who cannot stop taking a drug

c person that leads a company

d person that creates new products

excited: They were very excited.

a seeming larger than they really were

b highly respected

c feeling happy and enthusiastic

d looking unusual or foreign

pump: We need a new pump.

a equipment that forces liquid to move somewhere

b large, round vegetable with thick, orange skin

c place where you can keep food

d long boat with a flat bottom

slope: There was a big slope.

a animal that moves very slowly

b surface that is high at one end and low at the other

c lazy or dirty person

d long, narrow hole that you put something into

initiate: He tried to initiate it.

a steal

b start

c repair

d describe

upset: She upset him.

a made unhappy

b attracted his attention

c treated cruelly

d got nervous

greatly: He greatly admired her.

a from a distance

b in a polite way

c very much

d calmly

magic: It was a magic moment.

a special and exciting

b stupid and crazy

c strange and frightening

d short and fast

5K

object: I object to this.

a make it into a thing you can touch

b take a picture of high-quality

c say that you do not like something

d treat something like a tool or toy

oak: It was made of oak.

a type of flour made from a particular type of cereal

b wood of a large tree found in northern countries

c light material put in the top of a wine bottle

d small, white bubbles on the surface of a liquid

terrain: This is my terrain.

a particular type of land

b old kind of clothes

c dish made of small pieces of cooked meat

d flat area outside a house where you can sit

exploit: We could exploit this.

a make it burst with noise and force

b go around to find out what is there

c use for your advantage

d present something clearly and easy to understand

preliminary: This is preliminary.

a preventing something else from happening

b relating to the time before written records

c more important or better than others

d done in order to prepare for the main activity

rational: She tried to be rational.

a based on facts and not influenced by emotions

b making you feel less worried

c for political or social progress

d prepared for doing something

pencil: Can you give me a pencil?

a type of medicine that kills bacteria

b long, thin wooden object that you write with

c punishment for doing something against a rule

d coin used to pay in Britain

sack: She put it in the sack.

a large bag to carry things
b piece of furniture used for storing things
c container that is used to put waste in
d hollow space in something

aisle: She walked down the aisle.

a road on a small island
b footpath surrounding a large city
c passage between the lines of seats or goods
d street where people sell drugs

partially: I partially believe him.

a truly
b secretly
c not completely
d not in any way

6K

trophy: She got a trophy.

a metal instrument that you blow into
b piece of electric kitchen equipment
c medical condition that causes strong chest pains
d prize for winning a competition

broth: I like this broth.

a soup, usually made with meat
b building where people have sex for money
c something that is fastened onto clothes with a pin
d bread made with a special dough

forge: I told him to forge it.

a stop thinking about something
b decide not to be angry about something
c open it using physical strength
d make an illegal copy of something

wounded: He was wounded.

a worried
b injured
c persuaded
d surprised

skier: Are you a good skier?

a someone who slides over snow
b someone who flies airplanes
c someone who jumps from big heights
d someone who moves around on skates

jerk: Will it jerk backwards?

a slowly develop
b quickly turn
c suddenly move
d steadily change

transplant: They will use a transplant.

a vehicle that gets people from one place to another
b living thing that grows in the soil
c large factory where an industrial process happens
d operation in which an organ is put in a body

namely: I learned something, namely that apples are healthy.

a in particular
b in the name of
c in simple words
d in general

theft: There was a theft.

a book you could write in
b crime of stealing something
c building with a stage for plays
d long way from the top to the bottom

fortunate: I feel fortunate.

a strong
b comfortable
c attractive
d lucky

7K

altitude: What is the altitude here?

a age at which a person is allowed to drink alcohol
b difference in time zones
c one of two things that you can choose between
d height of something above sea level

tense: He was very tense.

a late
b nervous
c thankful
d happy

fashion: Can you fashion this for me?

a get it for me
b make it popular
c create it
d throw it away

ample: She had ample time.

a exactly enough

b more than enough

c not enough

d almost enough

poise: He had a lot of poise.

a money to buy unnecessary things

b bravery to do something difficult

c ability to behave in a calm way

d fear that something bad might happen

pathway: Follow the pathway.

a track that a person can walk along

b person in control of a group

c correct procedure of doing things

d large road vehicle for carrying goods

acquaintance: He is an acquaintance.

a someone who causes problems for you

b someone who has no place to live

c someone who has murdered several people

d someone who you know but not very well

decisive: She is not very decisive.

a making a choice quickly and easily

b making someone believe a lie

c making moral judgements about others

d making someone feel good about themselves

pasture: This is a nice pasture.

a area of land with grass where animals can feed

b generous gift for friends

c small cake that is made with pastry

d memory of happy times

thinker: She was a great thinker.

a someone who consumes a lot of alcohol

b someone who repairs something step by step

c someone who considers important subjects

d someone who helps very poor people

8K

secrecy: Secrecy is important.

a producing a substance from trees

b skill in dealing with people well

c being alone so that people cannot hear you

d not telling other people about it

graceful: She was graceful.

a behaving in a polite and pleasant way

b feeling or showing thanks

c making you feel guilty

d talking in a sad voice

appraisal: An appraisal is needed.

a person who is learning a job

b examination of something to judge how good it is

c improved version of a computer program

d official permission to do something

spelling: That is the correct spelling.

a way to write words

b way to bake bread

c way to argue with someone

d way to greet people

paralyze: I was paralysed.

a unable to move part of the body

b at a particular place at the same time

c caught in a strange situation

d too full of food

incorrect: This is incorrect.

a too big to measure

b different in colour

c unable to do work

d containing mistakes

reverse: I need to reverse.

a show more respect

b change something

c drive backwards

d write poetry

space: There was not much space.

a quality food

b fast speed

c loud noise

d empty area

precaution: This is just a precaution.

a something to prevent bad things in the future

b something you think will happen soon

c something used as an example

d something that comes before the main thing

sunrise: Can you see the sunrise?

a sun appearing in the morning

b sun going down in the evening

c sun at its highest point at noon

d sun being hidden by the moon

9K

musical: I liked the musical.

a sounds made by playing instruments

b group of people skilled in drama

c tunes recorded on a small disc

d film in which people sing and dance

immerse: I was so immersed.

a convinced the opposite was true

b completely involved in something

c shocked and very angry

d wanting to be successful

hierarchical: It was very hierarchical.

a arranged according to importance

b relating to events in the past

c extremely strange and funny

d using logical choice and reason

pathogen: She said it was a pathogen.

a situation that makes you feel sympathy

b part of a system that controls characteristics

c virus that can cause disease

d legal right to make a particular product

farewell: It was a sad farewell.

a hello

b thanks

c get well

d good bye

woo: You need to woo her.

a give a lot of attention

b surprise

c shock

d take seriously

generalization: This is a generalization.

a statement about the overall situation

b officer of very high rank in the army

c division between male and female

d group of people in a society who are the same age

famine: When was the last famine?

a full week of warm weather

b period when people do not have enough food

c day when schools are closed

d time when there is a lot of violence

pristine: This car is in pristine condition.

a very old

b very expensive

c very good

d very rare

floral: There was floral decoration.

a covering the entire surface

b very colourful and pretty

c made from flowers

d with cheap plastic

10K

cupboard: It is in the cupboard.

a container for tasty liquids

b box made of soft metal

c place to keep toys

d furniture with shelves inside

phenomenal: It was phenomenal.

a extremely successful

b impossible to understand

c making people believe things

d relating to the human body

brisk: She made a brisk move.

a big and ambitious

b quick and energetic

c small and weak

d slow and painful

stringent: We need to be more stringent.

a severe

b modest

c gentle

d brave

barber: Where can I find a barber?

a place where alcoholic drinks are sold

b someone whose job is to cut men's hair

c strong wire with short, sharp points on it

d meal that is prepared and eaten outdoors

lifting: You can do the lifting.

a put something in a higher position

b use a machine in tall buildings

c make somebody feel happy and joyful

d improve to look more attractive

inscribe: What could we inscribe here?

a give someone an idea for a book

b say something that makes people violent

c write words in a book or on an object

d draw something quickly and carelessly

high: I was on a high.

a top of a mountain

b feeling of excitement

c hot temperature

d having very good grades

scurry: He told me to scurry.

a try harder

b look at the night sky

c walk quickly

d use more spices

caffeine: I need caffeine.

a chemical that makes you feel more awake

b money to buy food

c feeling of being liked by people

d medicine used to reduce pain

11.6. Appendix F – Consent form population identification study (Chapter 4)

INFORMATION

As part of my PhD in the School of English, I am carrying out a study involving a test of vocabulary knowledge and the reading test you have just completed. I am going to analyse the scores of these tests, comparing them with each other to find out more about the relationship between reading ability and vocabulary knowledge.

I have approached you because I am interested in the lexical knowledge of learners of English. I would be very grateful if you agreed to take part.

You will now see a vocabulary test with 100 multiple-choice items. You have about 30 minutes to complete the test. Please answer only the questions where you are sure you know the answer. Do not guess.

You are free to withdraw from the study at any time. At every stage, your name will remain confidential. The data will be anonymized before the analysis and will be kept securely and used for academic purposes only.

Should you have any further queries about the study, please feel free to contact myself or my supervisor, Prof. Norbert Schmitt, who can be reached at norbert.schmitt@nottingham.ac.uk or by phone on +44 (0) 115 951 4847. You may also contact the Head of School, Prof. Josephine Guy, on +44 (0) 115 951 5921.

Benjamin Kremmel

benjamin.kremmel@nottingham.ac.uk

University of Nottingham
School of English
NG7 2RD
United Kingdom
Tel: +44 (0) 115 951 5900
http://www.nottingham.ac.uk/english/index.aspx

By starting this test you confirm that

- the purpose of the study has been explained to you and that you have understood it.
- you have had the opportunity to ask questions and they have been successfully answered.
- you understand that your participation in this study is voluntary and that you are free to not participate in the study, without giving a reason and without consequence
- you understand that all data are anonymous and that there will not be any connection between the personal information provided and the data.
- you understand that there are no known risks or hazards associated with participating in this study.
- you have read and understood the attached information and that agree to participate in this study.

11.7. Appendix G –IRT results from item piloting (Chapter 5)

```
TABLE 13.1 VKP_PILOT_VersionMASTER_Z VKP_PILOT_MASTER_RESULTS  Nov 28 10:08 2016
INPUT: 287 PERSON  435 ITEM  REPORTED: 287 PERSON  435 ITEM  2 CATS WINSTEPS 3.72.3
------------------------------------------------------------------------------------
PERSON: REAL SEP.: 3.27  REL.: .91 ... ITEM: REAL SEP.: 2.45  REL.: .86

         ITEM STATISTICS:  MEASURE ORDER

---------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|        |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM   |
|------------------------------------+----------+----------+-----------+-----------+--------|
|   419     12     64    4.40     .43| .91   -.2| .67   -.4|  .67   .63| 88.5  88.8| ITEM419|
|   427      9     39    3.97     .46| .78   -.8| .84   -.1|  .62   .53| 89.7  83.7| ITEM427|
|   380     15     65    3.92     .38| .91   -.3|1.00    .2|  .63   .61| 85.5  85.3| ITEM380|
|   432      9     38    3.90     .46|1.24    .9|1.87   1.5|  .38   .54| 78.9  83.8| ITEM432|
|   293     12     80    3.80     .34|1.02    .2| .90   -.1|  .34   .35| 86.3  86.1| ITEM293|
|   425     10     39    3.77     .44|1.28   1.1|1.17    .5|  .40   .54| 76.9  82.5| ITEM425|
|   422     17     64    3.64     .36| .96   -.1|1.49   1.2|  .58   .60| 83.6  83.0| ITEM422|
|   356     25     98    3.61     .27|1.10    .7|1.70   2.3|  .44   .51| 78.4  81.1| ITEM356|
|   360     26     99    3.53     .27|1.24   1.6|1.99   3.0|  .34   .52| 77.6  80.6| ITEM360|
|   396     27    101    3.47     .27|1.12    .8|1.42   1.5|  .42   .52| 76.0  80.3| ITEM396|
|   274     12     39    3.40     .42| .69  -1.4| .67   -.8|  .69   .54| 84.6  79.8| ITEM274|
|   401     28     96    3.33     .27| .92   -.5|1.21    .9|  .55   .52| 81.1  78.9| ITEM401|
|   379     19     63    3.32     .34|1.20   1.1|1.22    .7|  .51   .58| 73.3  80.5| ITEM379|
|   369     18     79    3.19     .29|1.13    .8|4.25   5.9|  .20   .38| 74.7  79.6| ITEM369|
|   372     19     80    3.15     .29| .95   -.3| .96   -.1|  .42   .39| 83.8  78.8| ITEM372|
|   413     19     78    3.14     .29| .87   -.9| .75  -1.0|  .52   .39| 80.8  78.3| ITEM413|
|   321     33     99    3.03     .25| .86  -1.1| .79  -1.0|  .60   .52| 79.6  76.6| ITEM321|
|   411     21     78    2.96     .28|1.08    .6|1.08    .4|  .32   .39| 73.1  76.6| ITEM411|
|   366     22     81    2.92     .27| .96   -.2| .93   -.2|  .42   .39| 80.2  76.4| ITEM366|
|   418     24     64    2.86     .31|1.19   1.3|1.24    .9|  .47   .56| 72.1  75.6| ITEM418|
|   332     23     80    2.84     .27|1.02    .2|1.13    .6|  .36   .39| 75.0  75.3| ITEM332|
|   388     15     38    2.81     .40|1.08    .5|1.10    .4|  .50   .54| 71.1  75.7| ITEM388|
|   373     24     81    2.77     .27| .90   -.8| .81   -.7|  .49   .39| 76.5  74.8| ITEM373|
|   249     37     96    2.74     .25|1.15   1.3|1.22   1.2|  .42   .51| 70.5  74.5| ITEM249|
|   367     24     78    2.70     .27| .83  -1.4| .79   -.8|  .54   .40| 83.3  74.4| ITEM367|
|   333     25     80    2.68     .27|1.18   1.4|2.99   5.7|  .19   .40| 67.5  73.9| ITEM333|
|   335     24     76    2.67     .27| .86  -1.1| .81   -.8|  .51   .40| 81.6  73.8| ITEM335|
|   203     40    100    2.64     .24| .84  -1.6| .80  -1.1|  .60   .51| 80.8  73.6| ITEM203|
|   329     26     80    2.63     .26|1.09    .7|1.20   1.1|  .31   .40| 71.3  73.1| ITEM329|
|   426     17     39    2.60     .38| .71  -1.9| .57  -1.5|  .70   .53| 84.6  74.1| ITEM426|
|   410     27     80    2.55     .26| .95   -.4| .88   -.5|  .45   .40| 73.8  72.5| ITEM410|
|   242     43     99    2.49     .24| .85  -1.6| .78  -1.3|  .59   .50| 79.6  72.1| ITEM242|
|   272     18     39    2.46     .38|1.29   1.7|1.18    .6|  .38   .52| 59.0  73.2| ITEM272|
|   312     18     39    2.46     .38| .89   -.6|1.02    .2|  .57   .52| 79.5  73.2| ITEM312|
|   400     43     97    2.45     .24|1.13   1.3|1.07    .5|  .43   .50| 62.5  72.0| ITEM400|
|   377     29     65    2.41     .30|1.00    .1| .98    .0|  .52   .53| 72.6  72.2| ITEM377|
|   285     44     99    2.39     .24| .80  -2.1| .70  -1.9|  .63   .50| 77.6  72.3| ITEM285|
|   390     18     38    2.38     .38|1.34   1.9|1.33   1.0|  .35   .53| 63.2  73.1| ITEM390|
|   384     29     64    2.37     .30|1.24   1.8|1.70   2.4|  .40   .53| 65.6  72.3| ITEM384|
|   371     30     81    2.37     .25| .89  -1.0| .82   -.9|  .50   .40| 79.0  70.9| ITEM371|
|   330     30     80    2.32     .25| .81  -2.0| .73  -1.5|  .57   .40| 76.3  70.5| ITEM330|
|   345     30     65    2.32     .30| .74  -2.3| .62  -1.7|  .64   .52| 80.6  71.7| ITEM345|
|   358     45     97    2.31     .24|1.06    .7| .99    .0|  .47   .50| 69.8  71.8| ITEM358|
|   323     46    100    2.30     .23| .95   -.5| .92   -.4|  .53   .50| 71.7  71.6| ITEM323|
|   281     47    101    2.28     .23| .99   -.1| .94   -.3|  .51   .50| 74.0  71.4| ITEM281|
|   431     19     38    2.27     .38| .92   -.5| .77   -.7|  .58   .52| 76.3  72.1| ITEM431|
|   167     48    100    2.22     .23|1.33   3.2|1.47   2.4|  .29   .49| 59.6  71.0| ITEM167|
|   404     46     94    2.19     .24| .80  -2.2| .74  -1.5|  .62   .50| 83.9  71.4| ITEM404|
|   189     32     65    2.15     .29| .81  -1.7|1.36   1.3|  .57   .51| 83.9  71.0| ITEM189|
|   391     20     38    2.14     .38| .75  -1.7| .65  -1.1|  .65   .51| 84.2  71.5| ITEM391|
|   338     32     64    2.12     .30| .95   -.4| .84   -.5|  .54   .51| 68.9  71.1| ITEM338|
|   201    137    287    2.12     .14| .98   -.4| .95   -.4|  .49   .48| 72.4  70.5| ITEM201|
|   238     20     39    2.11     .38|1.05    .4|1.02    .2|  .50   .52| 71.8  72.3| ITEM238|
|   291     35     80    2.05     .25| .98   -.2|1.02    .2|  .41   .40| 71.3  68.4| ITEM291|
|   236     21     40    2.05     .37| .90   -.6| .85   -.4|  .57   .51| 77.5  71.8| ITEM236|
|   319     50     99    2.05     .23|1.08    .9|1.02    .2|  .45   .49| 69.4  70.7| ITEM319|
|   308     21     39    2.03     .38|1.08    .6|1.32   1.0|  .44   .51| 61.5  71.5| ITEM308|
|   392     21     39    2.03     .38| .78  -1.6| .63  -1.1|  .64   .51| 82.1  71.5| ITEM392|
|   434     21     39    2.03     .38| .76  -1.8| .66  -1.0|  .64   .51| 82.1  71.5| ITEM434|
|   228     34     65    1.98     .29| .95   -.4| .93   -.2|  .51   .49| 75.8  70.7| ITEM228|
|   131     36     80    1.98     .25|1.09   1.0|1.11    .7|  .33   .41| 67.5  68.3| ITEM131|
|   250     37     80    1.90     .25| .96   -.5|1.73   3.6|  .40   .41| 76.3  68.0| ITEM250|
|   417     35     64    1.87     .29|1.04    .4| .95   -.1|  .48   .49| 63.9  70.5| ITEM417|
|   423     35     64    1.87     .29| .88  -1.1| .80   -.7|  .54   .49| 77.0  70.5| ITEM423|
|   368     38     80    1.85     .25|1.14   1.6|2.02   4.8|  .23   .41| 62.5  67.7| ITEM368|
|   344     36     65    1.81     .29| .87  -1.2| .75   -.9|  .55   .48| 72.6  70.4| ITEM344|
|   209     56    101    1.80     .23| .80  -2.3| .73  -1.5|  .59   .47| 85.0  71.0| ITEM209|
|   136     39     78    1.74     .25|1.13   1.5|1.08    .5|  .31   .40| 56.4  67.5| ITEM136|
|   297     37     65    1.72     .29|1.33   2.8|1.30   1.1|  .32   .48| 56.5  70.4| ITEM297|
|    424     37     64     1.69      .29| .91   -.8| .80   -.6|  .52    .47| 75.4  70.5| ITEM424|
|    218     41     80     1.65      .25|1.07    .8|1.11    .7|  .34    .40| 67.5  67.1| ITEM218|
|    234     24     40     1.64      .37| .87   -.9| .75   -.6|  .57    .49| 72.5  72.0| ITEM234|
|    310     23     37     1.64      .39|1.12    .8|1.03    .2|  .42    .48| 75.7  72.2| ITEM310|
|    359     58     98     1.63      .24| .87  -1.4| .79  -1.0|  .55    .47| 75.3  71.7| ITEM359|
|    212     42     81     1.63      .24| .91  -1.1|1.55   2.9|  .43    .40| 75.3  67.2| ITEM212|
|    270     24     39     1.61      .38|1.19   1.3|1.54   1.3|  .35    .48| 66.7  72.0| ITEM270|
|    271     24     39     1.61      .38| .92   -.5| .74   -.6|  .54    .48| 66.7  72.0| ITEM271|
|    395    165    286     1.60      .14| .94  -1.1| .85  -1.2|  .49    .45| 72.7  70.1| ITEM395|
|    327     43     80     1.57      .24| .99   -.1| .98   -.1|  .40    .39| 71.3  66.7| ITEM327|
|    357     59     97     1.56      .24| .66  -3.9| .55  -2.4|  .66    .46| 85.4  72.0| ITEM357|
|    266     39     64     1.50      .30|1.16   1.4|1.33   1.0|  .37    .46| 67.2  70.9| ITEM266|
|    196     25     40     1.50      .38|1.10    .7|1.52   1.2|  .40    .48| 72.5  72.7| ITEM196|
|    353     25     39     1.46      .38|1.19   1.2|1.08    .3|  .38    .47| 69.2  72.8| ITEM353|
|    386     25     39     1.46      .38| .88   -.8| .69   -.7|  .56    .47| 79.5  72.8| ITEM386|
|    245     61     96     1.39      .24| .83  -1.8| .70  -1.3|  .56    .44| 75.8  72.3| ITEM245|
|    221     41     65     1.38      .29|1.07    .6|1.07    .3|  .41    .45| 71.0  71.3| ITEM221|
|    260     41     65     1.38      .29| .92   -.7| .81   -.5|  .49    .45| 77.4  71.3| ITEM260|
|    299     41     65     1.38      .29|1.02    .2|1.05    .3|  .43    .45| 71.0  71.3| ITEM299|
|     46     64     99     1.33      .24|1.10   1.0|1.14    .7|  .38    .44| 66.3  72.7| ITEM046|
|    351     26     39     1.31      .39|1.02    .2| .85   -.2|  .47    .46| 74.4  73.9| ITEM351|
|    318     64     99     1.29      .24| .97   -.3| .83   -.6|  .48    .44| 74.5  72.7| ITEM318|
|    383     42     65     1.29      .30|1.17   1.4|1.08    .3|  .37    .44| 62.9  71.7| ITEM383|
|    208     64     98     1.27      .24| .90  -1.0| .79   -.8|  .51    .44| 74.2  73.0| ITEM208|
|    370     48     81     1.27      .25|1.13   1.4|1.51   2.6|  .26    .40| 54.3  67.3| ITEM370|
|    397     66    101     1.26      .24| .92   -.8| .81   -.8|  .49    .44| 77.0  72.8| ITEM397|
|     58     48     80     1.24      .25|1.28   2.8|1.48   2.4|  .14    .40| 58.8  67.7| ITEM058|
|    141     42     64     1.23      .30|1.18   1.5|1.17    .6|  .35    .44| 63.9  72.0| ITEM141|
|     55     48     79     1.23      .25| .96   -.4|1.25   1.3|  .40    .39| 68.4  67.6| ITEM055|
|    230     43     65     1.20      .30| .88  -1.0| .75   -.6|  .49    .43| 77.4  72.1| ITEM230|
|    110     27     40     1.20      .39| .85   -.9| .66   -.6|  .56    .46| 77.5  74.7| ITEM110|
|    407     49     80     1.18      .25|1.00    .0| .99    .0|  .40    .40| 65.0  68.1| ITEM407|
|    275     27     39     1.16      .40|1.12    .7| .88   -.1|  .41    .45| 64.1  75.3| ITEM275|
|    428     27     39     1.16      .40|1.07    .5|1.33    .8|  .38    .45| 79.5  75.3| ITEM428|
|    151     27     39     1.15      .40| .89   -.6| .77   -.3|  .52    .46| 79.5  75.5| ITEM151|
|    142     43     64     1.13      .30|1.24   1.9|1.65   1.6|  .30    .43| 65.6  72.4| ITEM142|
|    174     51     81     1.09      .25|1.01    .1| .93   -.3|  .40    .39| 72.8  68.8| ITEM174|
|     13     51     80     1.07      .25|1.26   2.6|1.66   2.9|  .12    .39| 58.8  69.0| ITEM013|
|    317    193    286     1.06      .14| .91  -1.5| .82  -1.2|  .47    .41| 75.9  72.6| ITEM317|
|    340     44     64     1.06      .31| .70  -2.5| .55  -1.2|  .57    .42| 80.3  73.3| ITEM340|
|    114     28     40     1.05      .39|1.10    .6|1.88   1.5|  .35    .45| 75.0  76.2| ITEM114|
|    337     45     65     1.02      .30| .91   -.6| .79   -.4|  .46    .42| 79.0  73.4| ITEM337|
|    215     52     81     1.02      .25|1.11   1.1|1.45   2.1|  .26    .39| 66.7  69.6| ITEM215|
|    365     52     81     1.02      .25|1.10   1.1|1.08    .5|  .31    .39| 61.7  69.6| ITEM365|
|    324     70    100      .98      .24| .99    .0| .88   -.3|  .43    .42| 77.8  74.3| ITEM324|
|    305     45     64      .96      .31| .80  -1.5| .70   -.7|  .51    .42| 77.0  74.1| ITEM305|
|    398     69     97      .95      .25| .92   -.7| .78   -.7|  .48    .42| 83.3  75.3| ITEM398|
|    182     46     65      .93      .31|1.26   1.8|1.13    .4|  .30    .41| 61.3  74.3| ITEM182|
|    362     71     99      .92      .25|1.28   2.3|1.35   1.2|  .23    .41| 68.4  75.2| ITEM362|
|    283     72    101      .91      .25|1.01    .1| .98    .0|  .40    .41| 77.0  75.0| ITEM283|
|    254     53     80      .90      .26| .94   -.5| .92   -.3|  .43    .39| 73.8  70.9| ITEM254|
|    217     53     80      .90      .26|1.00    .1| .94   -.2|  .39    .38| 66.3  70.8| ITEM217|
|    355    201    286      .90      .14|1.02    .3|1.13    .8|  .38    .40| 75.2  74.2| ITEM355|
|    120     72    100      .89      .25| .88  -1.1| .69  -1.0|  .50    .41| 74.7  75.5| ITEM120|
|    258     54     80      .85      .26|1.05    .5|1.17    .8|  .32    .39| 73.8  71.8| ITEM258|
|    192     28     38      .85      .42| .98    .0| .71   -.3|  .49    .45| 73.7  78.9| ITEM192|
|    364     72     99      .84      .25| .86  -1.3| .72   -.9|  .50    .40| 78.6  75.7| ITEM364|
|    300     47     65      .83      .31|1.19   1.3|2.11   2.1|  .25    .40| 75.8  75.3| ITEM300|
|    313     29     39      .83      .41| .92   -.3| .70   -.4|  .49    .43| 82.1  78.6| ITEM313|
|    385     29     39      .83      .41|1.20   1.0| .97    .2|  .34    .43| 71.8  78.6| ITEM385|
|    387     29     39      .83      .41| .97   -.1| .72   -.3|  .47    .43| 71.8  78.6| ITEM387|
|    393     29     39      .83      .41| .93   -.3| .91    .1|  .46    .43| 87.2  78.6| ITEM393|
|    292     54     80      .83      .26|1.00    .0| .94   -.2|  .39    .38| 76.3  71.7| ITEM292|
|    433     29     38      .75      .43| .92   -.3| .87    .0|  .46    .42| 81.6  80.1| ITEM433|
|    241    209    287      .74      .15| .89  -1.7| .73  -1.5|  .47    .39| 77.0  75.9| ITEM241|
|    259     48     65      .73      .32|1.04    .3| .95    .1|  .38    .39| 75.8  76.3| ITEM259|
|    264     48     65      .73      .32| .91   -.6| .80   -.3|  .44    .39| 79.0  76.3| ITEM264|
|    346     48     65      .73      .32| .98   -.1| .87   -.1|  .40    .39| 79.0  76.3| ITEM346|
|    322     72     97      .73      .26| .93   -.6| .77   -.6|  .45    .39| 78.1  76.9| ITEM322|
|    191     30     40      .73      .41|1.04    .2| .93    .1|  .41    .43| 80.0  79.1| ITEM191|
|    343     48     64      .70      .32| .84  -1.1| .81   -.3|  .45    .38| 82.0  76.9| ITEM343|
|    186     48     64      .68      .32| .91   -.5| .70   -.5|  .44    .39| 77.0  77.2| ITEM186|
|    148     48     64      .67      .32| .89   -.7| .72   -.5|  .45    .39| 80.3  77.1| ITEM148|
|    429     30     39      .66      .43| .95   -.2| .71   -.3|  .46    .41| 82.1  80.2| ITEM429|
|    375     48     64      .64      .32|1.08    .6| .99    .1|  .34    .38| 73.8  77.0| ITEM375|
|    382     49     65      .63      .32|1.01    .1| .88   -.1|  .38    .38| 74.2  77.3| ITEM382|
|    160    214    287      .63      .15|1.16   2.2|1.10    .6|  .29    .38| 70.7  77.1| ITEM160|
|    176     57     80      .62      .27| .95   -.4| .83   -.7|  .43    .37| 73.8  74.5| ITEM176|
|    273     30     38      .60      .44|1.16    .7| .89    .1|  .32    .39| 76.3  81.1| ITEM273|
|    267     48     63      .56      .33| .92   -.4| .89    .0|  .41    .38| 81.7  77.8| ITEM267|
|    243     76     99      .54      .26| .87  -1.0| .71   -.8|  .47    .38| 83.7  78.7| ITEM243|
|     52     58     79      .53      .28|1.26   1.8|1.60   2.0|  .11    .37| 72.2  76.1| ITEM052|
|    286     77    100      .53      .26|1.22   1.6|1.31    .9|  .24    .38| 73.7  78.8| ITEM286|
|    282     78    101      .53      .26| .96   -.2| .79   -.5|  .41    .38| 80.0  79.0| ITEM282|
|     23     50     65      .53      .33|1.30   1.8|2.16   1.9|  .17    .37| 74.2  78.3| ITEM023|
|    298     50     65      .53      .33| .91   -.5| .80   -.2|  .42    .37| 80.6  78.3| ITEM298|
|    381     50     65      .53      .33|1.20   1.2|1.23    .6|  .27    .37| 77.4  78.3| ITEM381|
|    315     30     38      .52      .45|1.00    .1| .85    .0|  .41    .41| 86.8  81.7| ITEM315|
|   227     49     63      .49      .34|1.05     .3| .92    .0|   .36    .38| 75.0  79.1| ITEM227|
|   144     49     63      .48      .34| .91    -.5| .70   -.4|   .43    .37| 81.7  79.0| ITEM144|
|   307     31     39      .47      .44| .88    -.4| .77   -.1|   .46    .40| 82.1  82.0| ITEM307|
|   205     75     96      .45      .27|1.12     .9|1.06    .3|   .31    .37| 75.8  79.8| ITEM205|
|   185     50     64      .43      .34|1.46    2.5|1.48    .9|   .14    .37| 67.2  79.1| ITEM185|
|   197     30     38      .42      .45| .97     .0| .73   -.1|   .44    .41| 78.9  81.6| ITEM197|
|   363     78     99      .42      .27|1.08     .6|1.05    .3|   .32    .37| 78.6  80.3| ITEM363|
|   336     59     79      .41      .28| .97    -.2|1.02    .2|   .38    .37| 75.9  77.1| ITEM336|
|   239     31     39      .39      .45| .86    -.5| .61   -.3|   .50    .41| 89.7  81.9| ITEM239|
|   204     80    101      .39      .27| .88    -.8| .69   -.7|   .45    .37| 83.0  80.6| ITEM204|
|   157     32     40      .36      .44|1.18     .8| .84    .0|   .34    .40| 77.5  82.2| ITEM157|
|   194     32     40      .36      .44| .89    -.4| .68   -.2|   .47    .40| 82.5  82.2| ITEM194|
|   232     32     40      .36      .44| .71   -1.2| .56   -.5|   .55    .40| 87.5  82.2| ITEM232|
|   135     60     79      .36      .29| .93    -.4| .87   -.4|   .42    .37| 81.0  78.2| ITEM135|
|   253     61     80      .32      .28|1.06     .5|1.06    .3|   .32    .36| 76.3  78.4| ITEM253|
|   295     60     78      .32      .29|1.15    1.0|1.40   1.3|   .15    .32| 75.6  78.0| ITEM295|
|   152     31     38      .32      .47|1.26    1.0|4.40   2.7|   .10    .40| 84.2  83.8| ITEM152|
|   405     62     81      .31      .28| .98    -.1| .84   -.5|   .40    .36| 77.8  78.6| ITEM405|
|   268     32     39      .26      .46| .88    -.4| .56   -.4|   .48    .38| 82.1  84.0| ITEM268|
|   347     32     39      .26      .46| .94    -.1| .70   -.1|   .43    .38| 82.1  84.0| ITEM347|
|   430     32     39      .26      .46| .91    -.2| .87    .1|   .42    .38| 87.2  84.0| ITEM430|
|   420     51     63      .23      .35|1.05     .3|1.35    .7|   .30    .35| 83.3  81.5| ITEM420|
|   198     32     39      .20      .47| .88    -.4| .92    .2|   .44    .39| 82.1  83.9| ITEM198|
|   170     62     79      .20      .30| .97    -.2| .98    .0|   .37    .35| 81.0  80.3| ITEM170|
|   223     53     65      .18      .35|1.02     .2| .93    .1|   .33    .34| 83.9  81.9| ITEM223|
|   229     53     65      .18      .35| .81   -1.0| .57   -.6|   .44    .34| 83.9  81.9| ITEM229|
|   303     53     65      .18      .35| .84    -.8| .81   -.1|   .41    .34| 87.1  81.9| ITEM303|
|   155     33     40      .16      .46| .81    -.7| .50   -.5|   .51    .38| 87.5  84.2| ITEM155|
|   156     33     40      .16      .46| .63   -1.5| .41   -.7|   .58    .38| 87.5  84.2| ITEM156|
|   143     51     62      .15      .37| .93    -.3| .83   -.1|   .38    .35| 81.4  82.6| ITEM143|
|   219     64     81      .15      .30| .89    -.7| .78   -.7|   .46    .36| 81.5  80.7| ITEM219|
|   124     82     99      .12      .29|1.08     .5| .86   -.2|   .31    .34| 81.6  83.6| ITEM124|
|   277    234    286      .12      .17|1.15    1.6|1.11    .5|   .26    .34| 80.9  82.8| ITEM277|
|   374     64     80      .08      .30| .96    -.2|1.02    .2|   .37    .36| 83.8  81.6| ITEM374|
|   320     83    100      .08      .29| .83   -1.1| .60   -.9|   .45    .34| 85.9  83.7| ITEM320|
|   409     64     80      .08      .30| .88    -.6| .72   -.8|   .46    .36| 83.8  81.6| ITEM409|
|   334     63     79      .07      .30| .95    -.3| .82   -.5|   .41    .35| 81.0  81.4| ITEM334|
|   216     64     80      .06      .30| .89    -.6| .81   -.5|   .44    .35| 82.5  81.6| ITEM216|
|   214     65     81      .06      .30|1.01     .1| .88   -.2|   .36    .35| 80.2  81.8| ITEM214|
|   328     65     81      .06      .30|1.14     .8|1.15    .6|   .23    .35| 80.2  81.8| ITEM328|
|   150     32     38      .06      .50| .85    -.4| .81    .1|   .44    .38| 89.5  85.8| ITEM150|
|   301     54     65      .05      .36| .72   -1.4| .46   -.9|   .47    .33| 87.1  83.2| ITEM301|
|   378     54     65      .05      .36| .81    -.9| .67   -.4|   .42    .33| 87.1  83.2| ITEM378|
|   352     33     39      .03      .49|1.48    1.5|5.04   2.8|  -.08    .36| 84.6  86.1| ITEM352|
|   394     33     39      .03      .49| .82    -.5| .59   -.3|   .47    .36| 89.7  86.1| ITEM394|
|   118     83     99      .01      .30|1.00     .0| .98    .1|   .33    .33| 86.7  84.5| ITEM118|
|   402     78     92      .01      .31|1.02     .2|1.23    .6|   .28    .32| 86.8  85.2| ITEM402|
|   171     64     79     -.01      .31|1.03     .3| .88   -.2|   .35    .35| 81.0  82.6| ITEM171|
|   179     64     79     -.03      .31| .92    -.4| .76   -.6|   .43    .35| 83.5  82.6| ITEM179|
|   128     66     81     -.04      .31|1.05     .3|1.17    .6|   .29    .35| 81.5  83.0| ITEM128|
|   207     84     99     -.07      .30|1.19    1.1|1.34    .8|   .19    .32| 83.7  85.4| ITEM207|
|   199     34     40     -.07      .49| .97     .0| .63   -.2|   .41    .36| 85.0  86.2| ITEM199|
|   237     34     40     -.07      .49| .84    -.5| .70   -.1|   .44    .36| 90.0  86.2| ITEM237|
|   222     55     65     -.08      .37| .97    -.1| .94    .1|   .33    .32| 85.5  84.7| ITEM222|
|   226     55     65     -.08      .37|1.16     .7| .95    .1|   .26    .32| 82.3  84.7| ITEM226|
|   376     55     65     -.08      .37| .97     .0| .75   -.3|   .34    .32| 85.5  84.7| ITEM376|
|    56     65     79     -.10      .32|1.08     .5|1.49   1.3|   .26    .35| 82.3  83.8| ITEM056|
|   133     67     81     -.13      .32|1.17     .9|1.67   1.6|   .16    .34| 82.7  84.1| ITEM133|
|   289     67     81     -.13      .32|1.01     .1|1.57   1.4|   .29    .34| 85.2  84.1| ITEM289|
|   325     86    100     -.14      .31|1.04     .3| .81   -.2|   .31    .31| 83.8  86.3| ITEM325|
|   220     67     80     -.14      .32| .92    -.4| .92   -.1|   .35    .29| 85.0  84.2| ITEM220|
|   403     82     95     -.17      .32|1.00     .1|1.19    .5|   .28    .31| 86.2  86.6| ITEM403|
|   278     87    101     -.19      .31|1.02     .2| .98    .1|   .29    .31| 87.0  86.5| ITEM278|
|   339     54     63     -.21      .39| .77    -.9| .50   -.8|   .43    .31| 86.7  85.8| ITEM339|
|   165     85     98     -.22      .32|1.10     .6| .80   -.2|   .29    .31| 85.6  87.1| ITEM165|
|   269     34     39     -.23      .53| .80    -.5| .58   -.2|   .45    .34| 89.7  88.1| ITEM269|
|   348     34     39     -.23      .53| .65   -1.1| .35   -.6|   .54    .34| 89.7  88.1| ITEM348|
|   389     34     39     -.23      .53| .99     .1| .70    .0|   .37    .34| 84.6  88.1| ITEM389|
|    65     56     65     -.23      .39|1.15     .7|1.60   1.0|   .23    .31| 83.9  86.2| ITEM065|
|   138     56     65     -.23      .39|1.08     .4| .83   -.1|   .28    .31| 87.1  86.2| ITEM138|
|   302     56     65     -.23      .39|1.13     .6|1.10    .4|   .24    .31| 83.9  86.2| ITEM302|
|   406     67     80     -.23      .33|1.00     .1|1.11    .4|   .33    .34| 83.8  85.1| ITEM406|
|   257     68     81     -.24      .33| .98     .0|1.20    .6|   .31    .34| 86.4  85.3| ITEM257|
|   161     87    100     -.26      .32| .83    -.8| .57   -.8|   .42    .30| 89.9  87.3| ITEM161|
|   126     87    100     -.27      .32|1.29    1.4|4.37   3.9|   .02    .31| 83.8  87.3| ITEM126|
|   117     87    100     -.27      .32|1.17     .9|1.70   1.3|   .16    .31| 85.9  87.3| ITEM117|
|   190     34     39     -.29      .53|1.20     .7|1.31    .6|   .21    .35| 89.7  88.0| ITEM190|
|   153     35     40     -.32      .52|1.23     .8|9.90   4.3|  -.03    .34| 85.0  88.3| ITEM153|
|    94     68     80     -.33      .34| .99     .0| .68   -.7|   .40    .34| 85.0  86.3| ITEM094|
|   175     67     78     -.34      .34| .85    -.6| .59  -1.0|   .44    .27| 87.2  86.2| ITEM175|
|   414     68     79     -.42      .35| .99     .1| .85   -.2|   .35    .34| 86.1  87.3| ITEM414|
|   408     68     79     -.42      .35|1.08     .4| .86   -.2|   .31    .34| 86.1  87.3| ITEM408|
|   168     88     99     -.46      .34|1.17     .8|2.32   1.9|   .12    .29| 87.8  88.9| ITEM168|
|   163     89    100     -.49      .34| .98     .0| .78   -.2|   .30    .29| 89.9  89.0| ITEM163|
|   247     88     99     -.50      .34| .92    -.3| .81   -.2|   .33    .28| 89.8  88.9| ITEM247|
|    44     88     98     -.52      .35|1.01     .1| .76   -.2|   .28    .27| 90.7  89.8| ITEM044|
|   262     57     64     -.53      .43|1.05     .3| .79   -.1|   .28    .28| 86.9  89.1| ITEM262|
|   421     57     64     -.54      .43| .89    -.3|1.33    .7|   .31    .28| 90.2  89.1| ITEM421|
|    240     34     38    -.56     .58| .87   -.2|1.25    .6|  .35    .33| 92.1  90.0| ITEM240|
|    181     58     65    -.57     .43|1.14    .6|1.28    .6|  .21    .28| 87.1  89.2| ITEM181|
|    415     58     65    -.57     .43| .87   -.4| .67   -.4|  .34    .28| 90.3  89.2| ITEM415|
|    361     88     98    -.58     .36| .91   -.3| .47   -.9|  .38    .28| 88.7  89.8| ITEM361|
|    111     36     40    -.62     .57|1.34    .9|1.03    .4|  .16    .31| 87.5  90.4| ITEM111|
|    122     91    100    -.74     .37|1.00    .1|1.24    .6|  .26    .26| 90.9  90.9| ITEM122|
|    288     72     80    -.76     .39|1.11    .5|1.26    .7|  .12    .24| 88.8  90.1| ITEM288|
|     21     59     65    -.76     .46|1.31   1.0|2.30   1.6|  .07    .26| 88.7  90.7| ITEM021|
|     29     59     65    -.76     .46|1.12    .5| .73   -.2|  .24    .26| 88.7  90.7| ITEM029|
|    139     59     65    -.76     .46| .82   -.5| .47   -.7|  .36    .26| 91.9  90.7| ITEM139|
|    265     59     65    -.76     .46| .90   -.2| .65   -.4|  .31    .26| 91.9  90.7| ITEM265|
|    399     90     98    -.82     .39| .99    .1| .51   -.8|  .30    .24| 91.8  91.8| ITEM399|
|    349     35     38    -.82     .65| .62   -.8| .23   -.6|  .51    .29| 94.7  92.3| ITEM349|
|     49     88     96    -.83     .39| .96   -.1| .75   -.2|  .28    .25| 91.6  91.6| ITEM049|
|    202     91     99    -.84     .39| .88   -.3| .57   -.6|  .34    .25| 91.8  91.9| ITEM202|
|    296     72     80    -.85     .40|1.11    .4| .76   -.3|  .29    .31| 90.0  90.8| ITEM296|
|    256     72     80    -.86     .40|1.07    .3|1.38    .8|  .22    .31| 90.0  90.8| ITEM256|
|    255     72     80    -.88     .40| .84   -.5| .60   -.7|  .44    .31| 92.5  90.8| ITEM255|
|    169     92    100    -.88     .39| .85   -.5| .56   -.6|  .35    .25| 91.9  92.0| ITEM169|
|    309     36     39    -.90     .65|1.17    .5| .65    .1|  .25    .28| 89.7  92.4| ITEM309|
|    354     36     39    -.90     .65| .88   -.1| .79    .2|  .33    .28| 94.9  92.4| ITEM354|
|    164     93    101    -.90     .39| .90   -.2| .54   -.7|  .33    .25| 92.0  92.0| ITEM164|
|    147     58     63    -.92     .50| .91   -.1| .76   -.1|  .28    .25| 93.3  91.9| ITEM147|
|     32     36     39    -.94     .65|1.43   1.0|9.90   5.5| -.21    .29| 89.7  92.3| ITEM032|
|    115     36     39    -.96     .65|1.39    .9|9.90   3.9| -.11    .29| 92.3  92.3| ITEM115|
|    193     36     39    -.96     .65| .64   -.7| .23   -.6|  .49    .29| 92.3  92.3| ITEM193|
|    231     36     39    -.96     .65| .88   -.1| .32   -.4|  .40    .29| 92.3  92.3| ITEM231|
|     78     35     38    -.96     .65| .96    .1| .75    .2|  .30    .28| 92.1  92.1| ITEM078|
|    342     58     63    -.96     .50|1.11    .4| .75   -.1|  .23    .25| 90.0  91.9| ITEM342|
|    261     59     64    -.96     .50| .95    .0| .60   -.4|  .28    .25| 93.4  92.0| ITEM261|
|    104     59     64    -.97     .50|1.14    .5| .69   -.2|  .22    .25| 90.2  92.0| ITEM104|
|     76     36     39    -.97     .65|1.21    .6|1.12    .5|  .18    .28| 92.3  92.3| ITEM076|
|     61     60     65    -.99     .49|1.16    .5|1.16    .5|  .17    .24| 90.3  92.1| ITEM061|
|    188     60     65    -.99     .49| .77   -.6| .38   -.8|  .36    .24| 93.5  92.1| ITEM188|
|    129     72     79   -1.00     .43| .91   -.2| .57   -.7|  .41    .30| 92.4  91.9| ITEM129|
|    119     91     98   -1.00     .41|1.06    .3| .86    .0|  .21    .24| 92.8  92.8| ITEM119|
|     82     92     99   -1.03     .41| .93   -.1|1.46    .8|  .24    .24| 92.9  92.9| ITEM082|
|    177     73     80   -1.05     .43|1.05    .3|1.35    .8|  .22    .30| 92.5  92.0| ITEM177|
|    294     73     80   -1.05     .43| .98    .0|1.20    .5|  .26    .30| 92.5  92.0| ITEM294|
|    210     92     99   -1.05     .41| .91   -.2|1.28    .6|  .26    .23| 92.9  92.9| ITEM210|
|    284     94    101   -1.06     .41|1.00    .1|5.35   3.6|  .18    .23| 93.0  93.0| ITEM284|
|    246     91     97   -1.18     .44|1.13    .5|1.47    .8|  .13    .22| 93.8  93.8| ITEM246|
|     60     60     64   -1.21     .55|1.01    .2|1.56    .9|  .16    .22| 95.1  93.5| ITEM060|
|    116    268    287   -1.21     .25|1.03    .2|1.80   1.8|  .18    .24| 93.3  93.5| ITEM116|
|    146     59     63   -1.22     .55| .85   -.2| .55   -.4|  .30    .23| 95.0  93.4| ITEM146|
|    172     74     80   -1.23     .46| .93   -.1| .66   -.4|  .37    .30| 93.8  93.1| ITEM172|
|    252     73     79   -1.24     .46| .79   -.5| .58   -.6|  .45    .29| 93.7  93.0| ITEM252|
|      6     95    101   -1.24     .44|1.09    .4|1.62   1.0|  .12    .22| 94.0  94.0| ITEM006|
|    279     95    101   -1.24     .44|1.08    .3|1.27    .6|  .14    .22| 94.0  94.0| ITEM279|
|    341     60     64   -1.24     .55| .81   -.3|1.04    .3|  .28    .22| 95.1  93.5| ITEM341|
|    331     75     81   -1.25     .46|1.04    .2|1.15    .4|  .24    .29| 93.8  93.2| ITEM331|
|    224     60     64   -1.25     .54| .77   -.5| .37   -.8|  .34    .22| 95.1  93.4| ITEM224|
|    187     61     65   -1.26     .54| .91   -.1| .74   -.1|  .25    .22| 95.2  93.5| ITEM187|
|    195     36     38   -1.38     .77| .73   -.3| .23   -.5|  .41    .25| 94.7  94.8| ITEM195|
|    350     37     39   -1.39     .77| .63   -.5| .16   -.7|  .45    .24| 94.9  94.9| ITEM350|
|    125     94     99   -1.42     .48| .93   -.1| .47   -.6|  .28    .20| 94.9  94.9| ITEM125|
|    149     37     39   -1.45     .77|1.28    .6|1.67    .9|  .05    .24| 94.9  94.9| ITEM149|
|    200     37     39   -1.45     .77| .78   -.2| .30   -.4|  .37    .24| 94.9  94.9| ITEM200|
|     57     74     79   -1.45     .50| .96    .0| .66   -.3|  .33    .29| 94.9  94.2| ITEM057|
|    112     37     39   -1.48     .77| .74   -.3| .24   -.5|  .40    .23| 94.9  94.9| ITEM112|
|     36     38     40   -1.48     .77|1.17    .5|1.28    .6|  .13    .24| 95.0  95.0| ITEM036|
|    109     38     40   -1.48     .77|1.27    .6|2.39   1.2|  .02    .24| 95.0  95.0| ITEM109|
|    154     38     40   -1.48     .77| .93    .1| .30   -.4|  .33    .24| 95.0  95.0| ITEM154|
|     26     61     64   -1.57     .62| .87   -.1|1.37    .7|  .21    .20| 95.1  95.1| ITEM026|
|     63     61     64   -1.58     .62| .86   -.1| .53   -.3|  .26    .20| 95.1  95.1| ITEM063|
|    225     62     65   -1.59     .62| .84   -.2|1.86   1.1|  .21    .20| 95.2  95.1| ITEM225|
|    306     62     65   -1.59     .62| .79   -.3| .40   -.5|  .29    .20| 95.2  95.1| ITEM306|
|    206     94     98   -1.65     .53|1.00    .1| .90    .1|  .19    .18| 95.9  95.9| ITEM206|
|    326     96    100   -1.69     .53| .96    .1| .49   -.5|  .24    .18| 96.0  96.0| ITEM326|
|     51     76     80   -1.74     .55| .92    .0| .90    .1|  .29    .27| 96.3  95.4| ITEM051|
|     11     77     81   -1.75     .55| .93    .0|1.58    .9|  .25    .27| 96.3  95.4| ITEM011|
|    244     94     97   -1.94     .60| .87   -.1|1.85   1.1|  .19    .16| 96.9  96.9| ITEM244|
|      5     94     97   -1.98     .60|1.05    .3| .93    .2|  .14    .16| 96.9  96.9| ITEM005|
|     43     96     99   -1.98     .60| .88   -.1| .88    .1|  .22    .16| 96.9  97.0| ITEM043|
|    184     62     64   -1.99     .74|1.07    .3| .51   -.2|  .17    .17| 96.7  96.7| ITEM184|
|     47     97    100   -2.00     .60|1.14    .4|9.90   5.2| -.12    .16| 97.0  97.0| ITEM047|
|     10     97    100   -2.00     .60|1.11    .4|1.58    .9|  .05    .16| 97.0  97.0| ITEM010|
|     62     61     63   -2.02     .74|1.17    .5|1.67    .9|  .05    .17| 96.7  96.7| ITEM062|
|     24     62     64   -2.03     .74|1.14    .4|1.71    .9|  .07    .17| 96.7  96.7| ITEM024|
|    101     62     64   -2.03     .74|1.03    .3| .51   -.2|  .18    .17| 96.7  96.7| ITEM101|
|    180     62     64   -2.03     .74|1.07    .3| .89    .3|  .14    .17| 96.7  96.7| ITEM180|
|    132     76     79   -2.04     .63| .69   -.5| .28   -.9|  .47    .26| 97.5  96.4| ITEM132|
|    145     62     64   -2.05     .74|1.12    .4| .80    .2|  .12    .16| 96.7  96.7| ITEM145|
|     97     62     64   -2.05     .74|1.01    .2| .39   -.4|  .20    .16| 96.7  96.7| ITEM097|
|    105     63     65   -2.05     .74|1.15    .4|2.45   1.3|  .04    .16| 96.8  96.8| ITEM105|
|    130     77     80   -2.05     .63| .65   -.6| .18  -1.2|  .51    .25| 97.5  96.4| ITEM130|
|     79    278    287   -2.08     .35| .92   -.2| .63   -.7|  .23    .18| 97.2  96.9| ITEM079|
|    134     77     80   -2.08     .63|1.14    .4| .79    .0|  .22    .25| 95.0  96.4| ITEM134|
|    96     78     81   -2.09     .63| .81   -.2| .44   -.5|  .38   .25| 97.5  96.5| ITEM096|
|   178     78     81   -2.09     .63| .73   -.4|1.19    .5|  .36   .25| 97.5  96.5| ITEM178|
|   290     78     81   -2.09     .63|1.23    .6| .79    .0|  .17   .25| 95.1  96.5| ITEM290|
|   311     38     39   -2.18    1.05|1.15    .5|1.20    .6|  .07   .18| 97.4  97.4| ITEM311|
|   314     38     39   -2.18    1.05|1.13    .4| .82    .3|  .12   .18| 97.4  97.4| ITEM314|
|   316     38     39   -2.18    1.05|1.06    .4| .41   -.1|  .19   .18| 97.4  97.4| ITEM316|
|   158     37     38   -2.20    1.05|1.06    .4| .43    .0|  .19   .18| 97.4  97.4| ITEM158|
|    74     37     38   -2.22    1.05|1.06    .4| .44    .0|  .19   .18| 97.4  97.4| ITEM074|
|   233     38     39   -2.23    1.05|1.04    .3| .37   -.1|  .21   .18| 97.4  97.4| ITEM233|
|    75     38     39   -2.25    1.05| .72    .0| .11   -.6|  .34   .17| 97.4  97.4| ITEM075|
|   159     38     39   -2.25    1.05|1.06    .4| .44    .0|  .18   .17| 97.4  97.4| ITEM159|
|    35     39     40   -2.26    1.05| .72    .0| .11   -.6|  .34   .17| 97.5  97.5| ITEM035|
|    38     39     40   -2.26    1.05| .94    .2| .22   -.4|  .26   .17| 97.5  97.5| ITEM038|
|   108     39     40   -2.26    1.05|1.16    .5|2.12   1.1|  .01   .17| 97.5  97.5| ITEM108|
|   235     39     40   -2.26    1.05|1.15    .5|1.51    .8|  .05   .17| 97.5  97.5| ITEM235|
|   248     95     97   -2.36     .73|1.05    .3|2.87   1.6|  .04   .13| 97.9  97.9| ITEM248|
|    83     95     97   -2.40     .73|1.01    .2| .49   -.3|  .16   .13| 97.9  97.9| ITEM083|
|   162     98    100   -2.42     .73| .96    .2| .42   -.4|  .19   .13| 98.0  98.0| ITEM162|
|    81     98    100   -2.44     .73|1.10    .4|4.35   2.3| -.07   .13| 98.0  98.0| ITEM081|
|   280     99    101   -2.44     .73|1.01    .2| .46   -.4|  .16   .13| 98.0  98.0| ITEM280|
|    90     77     79   -2.53     .76| .76   -.2| .50   -.2|  .35   .23| 97.5  97.5| ITEM090|
|    54     78     80   -2.53     .76| .78   -.1|1.77    .9|  .25   .23| 97.5  97.5| ITEM054|
|    18     77     79   -2.55     .76| .77   -.2| .69    .0|  .32   .22| 97.5  97.5| ITEM018|
|    17     77     79   -2.55     .76|1.10    .4| .63    .0|  .23   .22| 97.5  97.5| ITEM017|
|    92     79     81   -2.57     .76| .61   -.5| .10  -1.2|  .49   .22| 97.5  97.5| ITEM092|
|   213     79     81   -2.57     .76| .74   -.2| .30   -.6|  .38   .22| 97.5  97.5| ITEM213|
|   251     79     81   -2.57     .76| .61   -.5| .10  -1.2|  .49   .22| 97.5  97.5| ITEM251|
|     1    282    287   -2.72     .47| .87   -.2|2.81   2.2|  .17   .14| 98.2  98.2| ITEM001|
|    64     63     64   -2.79    1.03| .73    .0| .09   -.8|  .26   .12| 98.4  98.4| ITEM064|
|   304     63     64   -2.80    1.03| .73    .0| .09   -.8|  .25   .12| 98.4  98.4| ITEM304|
|   263     64     65   -2.80    1.03| .98    .3| .27   -.4|  .17   .12| 98.4  98.4| ITEM263|
|    37     38     38   -3.09    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM037|
|     8     98     99   -3.09    1.02|1.03    .4| .86    .3|  .06   .09| 99.0  99.0| ITEM008|
|   121     96     97   -3.11    1.02| .99    .3| .28   -.5|  .15   .10| 99.0  99.0| ITEM121|
|    45     99    100   -3.14    1.02| .99    .3| .28   -.5|  .14   .09| 99.0  99.0| ITEM045|
|    85     98     99   -3.14    1.02|1.05    .4|1.90   1.0|  .01   .09| 99.0  99.0| ITEM085|
|    41     99    100   -3.14    1.02|1.06    .4|9.90   4.7| -.15   .09| 99.0  99.0| ITEM041|
|    80     99    100   -3.14    1.02|1.05    .4|2.42   1.3| -.01   .09| 99.0  99.0| ITEM080|
|   166     99    100   -3.15    1.02|1.06    .4|9.90   6.8| -.21   .09| 99.0  99.0| ITEM166|
|   123     99    100   -3.16    1.02|1.03    .4| .68    .1|  .08   .09| 99.0  99.0| ITEM123|
|    87     99    100   -3.16    1.02| .90    .2| .13   -.9|  .20   .09| 99.0  99.0| ITEM087|
|    19     78     79   -3.25    1.05| .57   -.2| .04  -1.1|  .43   .17| 98.7  98.7| ITEM019|
|   412     74     75   -3.26    1.05|1.22    .5|7.02   2.5| -.13   .19| 98.7  98.7| ITEM412|
|    53     78     79   -3.30    1.05| .59   -.2| .04  -1.0|  .43   .18| 98.7  98.7| ITEM053|
|    14     75     76   -3.30    1.05| .59   -.2| .04  -1.0|  .42   .18| 98.7  98.7| ITEM014|
|   127     77     78   -3.31    1.05| .59   -.2| .04  -1.0|  .42   .18| 98.7  98.7| ITEM127|
|    15     78     79   -3.33    1.05| .60   -.2| .04  -1.0|  .42   .18| 98.7  98.7| ITEM015|
|    59     79     80   -3.34    1.05|1.16    .5| .76    .2|  .10   .17| 98.8  98.7| ITEM059|
|    91     79     80   -3.34    1.05| .60   -.2| .04  -1.0|  .42   .17| 98.8  98.7| ITEM091|
|    50     79     80   -3.34    1.05| .60   -.2| .04  -1.0|  .42   .17| 98.8  98.7| ITEM050|
|   137     78     79   -3.34    1.05| .60   -.2| .04  -1.0|  .42   .17| 98.7  98.7| ITEM137|
|    93     79     80   -3.34    1.05| .60   -.2| .04  -1.0|  .42   .17| 98.8  98.7| ITEM093|
|   173     79     80   -3.34    1.05|1.18    .5|1.44    .7|  .03   .17| 98.8  98.7| ITEM173|
|    88     80     81   -3.35    1.05|1.02    .3| .15   -.6|  .27   .17| 98.8  98.8| ITEM088|
|    95     80     81   -3.35    1.05|1.17    .5|1.44    .7|  .03   .17| 98.8  98.8| ITEM095|
|   211     80     81   -3.35    1.05| .60   -.2| .04  -1.1|  .42   .17| 98.8  98.8| ITEM211|
|   287     80     81   -3.35    1.05| .60   -.2| .04  -1.1|  .42   .17| 98.8  98.8| ITEM287|
|   276     39     39   -3.45    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM276|
|    31     37     37   -3.45    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM031|
|    33     39     39   -3.49    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM033|
|    72     38     38   -3.50    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM072|
|    70     39     39   -3.50    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM070|
|    73     39     39   -3.51    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM073|
|   435     39     39   -3.52    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM435|
|    34     40     40   -3.53    1.85|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM034|
|    39     40     40   -3.53    1.85|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM039|
|    71     40     40   -3.53    1.85|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM071|
|    77     40     40   -3.53    1.85|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM077|
|   106     40     40   -3.53    1.85|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM106|
|   107     40     40   -3.53    1.85|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM107|
|   113     40     40   -3.53    1.85|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM113|
|    69     64     64   -3.82    1.83|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM069|
|   183     63     63   -3.98    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM183|
|    25     63     63   -3.98    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM025|
|    28     63     63   -4.02    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM028|
|    27     64     64   -4.03    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM027|
|    30     64     64   -4.03    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM030|
|    22     64     64   -4.03    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM022|
|   102     64     64   -4.03    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM102|
|    98     64     64   -4.04    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM098|
|    99     64     64   -4.04    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM099|
|   100     64     64   -4.04    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM100|
|    66     65     65   -4.05    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM066|
|    67     65     65   -4.05    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM067|
|    68     65     65   -4.05    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM068|
|   103     65     65   -4.05    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM103|
|   140     65     65   -4.05    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM140|
|   416     65     65   -4.05    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM416|
|    89     80     80    -4.25    1.83|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM089|
|    42     97     97    -4.25    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM042|
|     2     95     95    -4.35    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM002|
|     9     99     99    -4.36    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM009|
|    48     99     99    -4.36    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM048|
|    86     99     99    -4.37    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM086|
|     3     99     99    -4.37    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM003|
|     4     99     99    -4.37    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM004|
|    84     99     99    -4.38    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM084|
|     7    100    100    -4.38    1.84|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM007|
|    40    286    287    -4.41    1.01| .81    .1| .03  -2.6|  .18   .07| 99.6  99.6| ITEM040|
|    16     78     78    -4.58    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM016|
|    20     79     79    -4.61    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM020|
|    12     80     80    -4.62    1.86|       MINIMUM MEASURE|  .00   .00|100.0 100.0| ITEM012|
|-------------------------------------+----------+----------+-----------+-----------+--------|
| MEAN    59.4   75.3     -.41     .61| .98    .0|1.13    .2|           | 84.5  84.3|        |
| S.D.    35.9   40.2     2.07     .48| .17    .9|1.32   1.2|           | 10.6   9.9|        |
----------------------------------------------------------------------------------------------
```

11.8. Appendix H – Target words in item piloting (Chapter 5)

#index	# band	#target
1	500	speak
2	500	face
3	500	read
4	500	paper
5	500	stand
6	500	add
7	500	office
8	500	spend
9	500	door
10	500	health
11	500	person
12	500	art
13	500	different
14	500	war
15	500	history
16	500	party
17	500	grow
18	500	window
19	500	open
20	500	body
21	500	morning
22	500	walk
23	500	attention
24	500	low
25	500	win
26	500	research
27	500	girl
28	500	guy
29	500	early
30	500	food
31	500	line
32	500	air
33	500	teacher
34	500	force
35	500	offer
36	500	education
37	500	remember
38	500	foot
39	500	boy
40	1,000	sound
41	1,000	enjoy
42	1,000	network
43	1,000	legal
44	1,000	religious
45	1,000	cold
46	1,000	form
47	1,000	science
48	1,000	green
49	1,000	memory
50	1,000	card
51	1,000	seat
52	1,000	cell
53	1,000	sign
54	1,000	rich
55	1,000	trial
56	1,000	expert
57	1,000	spring
58	1,000	firm
59	1,000	radio
60	1,000	visit
61	1,000	management
62	1,000	care
63	1,000	avoid
64	1,000	imagine
65	1,000	huge
66	1,000	ball
67	1,000	finish
68	1,000	talk
69	1,000	garden
70	1,000	impact
71	1,000	bird
72	1,000	charge
73	1,000	popular
74	1,000	traditional
75	1,000	direction
76	1,000	weapon
77	1,000	kitchen
78	1,000	contain
79	1,500	suit
80	1,500	bus
81	1,500	growing
82	1,500	blow
83	1,500	construction
84	1,500	rain
85	1,500	destroy
86	1,500	cook
87	1,500	charge
88	1,500	connection
89	1,500	burn
90	1,500	shoe
91	1,500	photo

92	1,500	view
93	1,500	farmer
94	1,500	leaf
95	1,500	committee
96	1,500	lip
97	1,500	pair
98	1,500	smile
99	1,500	chicken
100	1,500	clothes
101	1,500	quiet
102	1,500	climb
103	1,500	promise
104	1,500	empty
105	1,500	complete
106	1,500	drive
107	1,500	circle
108	1,500	bone
109	1,500	active
110	1,500	extend
111	1,500	tape
112	1,500	combine
113	1,500	wine
114	1,500	below
115	1,500	cool
116	2,000	totally
117	2,000	hero
118	2,000	industrial
119	2,000	cloud
120	2,000	stretch
121	2,000	winner
122	2,000	volume
123	2,000	travel
124	2,000	seed
125	2,000	surprised
126	2,000	rest
127	2,000	fashion
128	2,000	pepper
129	2,000	busy
130	2,000	separate
131	2,000	intervention
132	2,000	copy
133	2,000	tip
134	2,000	cheap
135	2,000	cite
136	2,000	welfare
137	2,000	vegetable
138	2,000	dish
139	2,000	improvement
140	2,000	beach
141	2,000	gray
142	2,000	opening
143	2,000	divide
144	2,000	initial
145	2,000	terrible
146	2,000	oppose
147	2,000	route
148	2,000	contemporary
149	2,000	multiple
150	2,000	essential
151	2,000	question
152	2,000	league
153	2,000	careful
154	2,000	criminal
155	2,000	core
156	2,000	upper
157	2,000	rush
158	2,000	specifically
159	2,000	tired
160	2,500	shape
161	2,500	relative
162	2,500	educator
163	2,500	belt
164	2,500	immigration
165	2,500	teaspoon
166	2,500	birthday
167	2,500	implication
168	2,500	perfectly
169	2,500	coast
170	2,500	supporter
171	2,500	accompany
172	2,500	silver
173	2,500	teenager
174	2,500	recognition
175	2,500	retirement
176	2,500	recovery
177	2,500	flag
178	2,500	watch
179	2,500	whisper
180	2,500	gentleman
181	2,500	corn
182	2,500	inner
183	2,500	moon
184	2,500	junior
185	2,500	swing
186	2,500	throat
187	2,500	salary
188	2,500	observer
189	2,500	due

190	2,500	straight
191	2,500	publication
192	2,500	crop
193	2,500	pretty
194	2,500	permanent
195	2,500	plant
196	2,500	phenomenon
197	2,500	anxiety
198	2,500	literally
199	2,500	resist
200	2,500	wet
201	3,000	vessel
202	3,000	storage
203	3,000	flee
204	3,000	leather
205	3,000	distribute
206	3,000	ill
207	3,000	evolution
208	3,000	shelf
209	3,000	tribe
210	3,000	can
211	3,000	girlfriend
212	3,000	lawn
213	3,000	assistant
214	3,000	council
215	3,000	wisdom
216	3,000	vulnerable
217	3,000	garlic
218	3,000	instance
219	3,000	poetry
220	3,000	celebrity
221	3,000	gradually
222	3,000	stability
223	3,000	doubt
224	3,000	fantasy
225	3,000	scared
226	3,000	guide
227	3,000	plot
228	3,000	framework
229	3,000	gesture
230	3,000	ongoing
231	3,000	psychology
232	3,000	counselor
233	3,000	since
234	3,000	witness
235	3,000	chapter
236	3,000	fellow
237	3,000	divorce
238	3,000	resemble
239	3,000	pipe
240	3,000	athletic
241	4,000	sweat
242	4,000	undermine
243	4,000	outer
244	4,000	drunk
245	4,000	survey
246	4,000	research
247	4,000	separation
248	4,000	traditionally
249	4,000	ballot
250	4,000	stuff
251	4,000	intelligent
252	4,000	govern
253	4,000	driving
254	4,000	rhetoric
255	4,000	convinced
256	4,000	vitamin
257	4,000	enthusiasm
258	4,000	accommodate
259	4,000	wilderness
260	4,000	praise
261	4,000	injure
262	4,000	endless
263	4,000	pause
264	4,000	mandate
265	4,000	excuse
266	4,000	respectively
267	4,000	chaos
268	4,000	uncertainty
269	4,000	mechanical
270	4,000	format
271	4,000	canvas
272	4,000	profound
273	4,000	lobby
274	4,000	trait
275	4,000	currency
276	4,000	apologize
277	5,000	trouble
278	5,000	accelerate
279	5,000	happily
280	5,000	dancing
281	5,000	enact
282	5,000	removal
283	5,000	autonomy
284	5,000	disturb
285	5,000	thread
286	5,000	landmark
287	5,000	unhappy

288	5,000	privately
289	5,000	fraction
290	5,000	tourism
291	5,000	offender
292	5,000	distinctive
293	5,000	threshold
294	5,000	calm
295	5,000	suite
296	5,000	routinely
297	5,000	remark
298	5,000	regulator
299	5,000	straw
300	5,000	theological
301	5,000	fragile
302	5,000	exhaust
303	5,000	globe
304	5,000	chemistry
305	5,000	objection
306	5,000	old-fashioned
307	5,000	crowded
308	5,000	blast
309	5,000	circle
310	5,000	prevail
311	5,000	overnight
312	5,000	denial
313	5,000	fragment
314	5,000	headache
315	5,000	rental
316	5,000	fantastic
317	6,000	assurance
318	6,000	spark
319	6,000	chop
320	6,000	competing
321	6,000	mob
322	6,000	spare
323	6,000	weep
324	6,000	consultation
325	6,000	liquor
326	6,000	dioxide
327	6,000	accountable
328	6,000	affirm
329	6,000	pace
330	6,000	sip
331	6,000	sadly
332	6,000	span
333	6,000	emergence
334	6,000	lifelong
335	6,000	linger
336	6,000	applaud
337	6,000	stabilize
338	6,000	fold
339	6,000	cube
340	6,000	harbor
341	6,000	calm
342	6,000	terminal
343	6,000	embassy
344	6,000	preacher
345	6,000	dim
346	6,000	injection
347	6,000	antique
348	6,000	plantation
349	6,000	predictable
350	6,000	sunset
351	6,000	presume
352	6,000	x-ray
353	6,000	excess
354	6,000	empty
355	8,000	obesity
356	8,000	affluent
357	8,000	cozy
358	8,000	harbor
359	8,000	takeover
360	8,000	exacerbate
361	8,000	embarrass
362	8,000	milky
363	8,000	realism
364	8,000	knight
365	8,000	tangible
366	8,000	feat
367	8,000	groan
368	8,000	militant
369	8,000	dwell
370	8,000	forecast
371	8,000	razor
372	8,000	lurk
373	8,000	hay
374	8,000	spinach
375	8,000	plug
376	8,000	niece
377	8,000	swiftly
378	8,000	terminate
379	8,000	huddle
380	8,000	strap
381	8,000	tactical
382	8,000	space
383	8,000	attic
384	8,000	constellation
385	8,000	beetle

386	8,000	plague
387	8,000	populate
388	8,000	maneuver
389	8,000	pearl
390	8,000	probation
391	8,000	wreck
392	8,000	smack
393	8,000	abusive
394	8,000	civilized
395	10,000	caption
396	10,000	binding
397	10,000	devastation
398	10,000	healer
399	10,000	safeguard
400	10,000	larva
401	10,000	blaze
402	10,000	rapper
403	10,000	coordinate
404	10,000	blur
405	10,000	insulin
406	10,000	midday
407	10,000	interdisciplinary
408	10,000	barber
409	10,000	donkey
410	10,000	fallout
411	10,000	heed
412	10,000	last-minute
413	10,000	scam
414	10,000	malaria
415	10,000	horrific
416	10,000	unsafe
417	10,000	avoidance
418	10,000	liken
419	10,000	scant
420	10,000	allergic
421	10,000	licensed
422	10,000	lurch
423	10,000	comb
424	10,000	gamble
425	10,000	brisk
426	10,000	bounty
427	10,000	cramped
428	10,000	authoritative
429	10,000	scar
430	10,000	rocker
431	10,000	irritation
432	10,000	ostensibly
433	10,000	blindness
434	10,000	flea
435	1,000	mouth

11.9. Appendix I – Consent form for item piloting and algorithm study (Chapters 5 & 6)

INFORMATION

As part of my PhD in the School of English, I am carrying out a study involving different vocabulary test items. I am going to analyse the scores of these tests to analyse whether they are clear and functioning as intended.

I have approached you because I am interested in the lexical knowledge of speakers of English. I would be very grateful if you agreed to take part.

I will now give you a vocabulary test with about 100 items. Your knowledge of the words in this test will be assessed using a Multiple Choice test format. It will take you about 30-45 minutes to complete the test. Please answer only the questions where you are sure you know the answer. Do not guess.

You are free to withdraw from the study before starting the online test or to exit the online test at any time. Please note, because the survey data will be anonymous, it will not be possible to withdraw from the study after you have completed the test because your data will not be able to be identified. At every stage, your identity will thus remain confidential and any data will be kept securely and used for academic purposes only.

Should you have any further queries about the study, please feel free to contact myself or my supervisor, Prof. Norbert Schmitt, who can be reached at norbert.schmitt@nottingham.ac.uk or by phone on +44 (0) 115 951 4847. You may also contact the Head of School, Prof. Josephine Guy, on +44 (0) 115 951 5921.

Benjamin Kremmel

benjamin.kremmel@nottingham.ac.uk

University of Nottingham
School of English
NG7 2RD
United Kingdom
Tel: +44 (0) 115 951 5900
http://www.nottingham.ac.uk/english/index.aspx

By starting this test you confirm that

- the purpose of the study has been explained to you and that you have understood it.
- you have had the opportunity to ask questions and they have been successfully answered.
- you understand that your participation in this study is voluntary and that you are free to not participate in the study, without giving a reason and without consequence
- you understand that all data are anonymous and that there will not be any connection between the personal information provided and the data.
- you understand that there are no known risks or hazards associated with participating in this study.
- you have read and understood the attached information and that agree to participate in this study.

11.10. Appendix J – Vocabulary Knowledge Profiler

To view the beta version of the test, please go to:

http://vkp.benjaminkremmel.com

11.11. Appendix K – Consent form for Reading relationship study (Chapter 7)

The following text was displayed one the first page of the test.

> *I confirm that: (a) I am over 16, (b) I have understood the purpose of this study, (c) that all data are anonymous and that there will not be any connection between the personal information provided and the data, (d) there are no known risks or hazards associated participating in this study. By starting the tests, I agree that my answers, which I have given voluntarily, can be used anonymously for research purposes.*