DIAGNOSTIC TESTS OF ENGLISH VOCABULARY LEARNING PROFICIENCY: GUESSING FROM CONTEXT AND KNOWLEDGE OF WORD PARTS

By

Yosuke Sasao

A thesis

submitted to the Victoria University of Wellington

in fulfilment of the requirements for the degree of

Doctor of Philosophy

in Applied Linguistics

Victoria University of Wellington

2013

ABSTRACT

This thesis looked at the creation and validation of two tests that measure how efficiently English words are learned. Previous studies have created and validated a number of tests that measure the size (how many words are known) and the depth (how well a word is known) of vocabulary knowledge; however, existing vocabulary tests do not indicate how learners can become proficient in vocabulary learning. This research was one of the first attempts to create such tests. A guessing-from-context test (GCT) and a word part test (WPT) were created, because the skill of guessing from context and word part knowledge are teachable and are the most frequently used strategies for dealing with unknown words.

The GCT consisted of the following three sections: identifying the part of speech of an unknown word, finding the contextual clue that helps guess its meaning, and deriving the unknown word's meaning. Each of these three sections was designed to measure each of the important steps in guessing from context that was identified by previous studies. The test was validated using Rasch analysis through data from 428 Japanese learners of English. The results indicated that the GCT is a highly valid and reliable measure of the skill of guessing from context in terms of eight aspects of construct validity (content, substantial, structural, generalizability, external, consequential, responsiveness, and interpretability). Based on the results, two new equivalent forms were created in order to allow a pre- and post-test design where researchers and teachers can investigate learners' development of the skill of guessing from context.

The WPT measured 118 word parts that were selected based on frequency data in

the British National Corpus. It consisted of the following three sections: form (recognition of written word parts), meaning (knowledge of their meanings), and use (knowledge of their syntactic properties). These three sections were designed to measure the important aspects of word part knowledge that were identified by previous studies. The WPT was validated using Rasch analysis through data from 440 Japanese learners of English and 1,348 people with various native languages. The results indicated that the WPT is a highly valid and reliable measure of word part knowledge in terms of the eight aspects of construct validity mentioned above. As with the GCT, two new equivalent forms were created in order to allow a pre- and post-test design. For more practical use of the test, the Word Part Levels Test (WPLT) was created by classifying the 118 word parts into three different levels of difficulty. This may allow teachers to quickly examine whether their students need to work on easy or difficult word parts and which aspects of word part knowledge need to be learned. Taken as a whole, the GCT and the WPT are useful measures both for research and practical purposes.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my primary supervisor, Stuart Webb, for his keen insight and constructive criticism throughout my research. Without his generous support and direction, my research would not have progressed this far. I am also grateful to my secondary supervisor, Paul Nation, for his encouragement and insightful comments. I was extremely honoured to have Anna Siyanova, John Read, and Tom Cobb as thesis examiners.

I would like to thank Dalice Sim and Yuichi Hirose for their expert advice on statistical analyses. My thanks are also due to Laurie Bauer for his invaluable comments on the word part test. I gratefully acknowledge helpful discussions with graduate students at Victoria University of Wellington and Kyoto University. I have especially benefited from discussions with Mike Rodgers and Tatsuya Nakata. I am also deeply grateful to Myq Larson for making a web-based word part test available for my study.

My special thanks go to Akira Tajino who inspired me to do research into vocabulary acquisition. He taught me important skills for completing a doctoral thesis. I am also indebted to David Dalsky, Kazuyo Murata, Kenji Tani, Mariko Abe, Noriko Kurihara, and Sayako Maswana, for allowing me into their classes and helping me to collect data.

I gratefully acknowledge the financial support from Victoria University of Wellington in the form of a Victoria PhD Scholarship and a Faculty Research Grant.

I wish to express my gratitude to my family, Takeshi, Eiko, Yoshiko, Kanji, Hisami, and Keita, for their warm-hearted support during my research. I also wish to say “thank you” to my children, Kotaro, Kenjiro, and Konoka, whose smiles have been a great support to me. Finally, my deepest appreciation goes to my wife, Etsuko, whose patient love enabled me to complete this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Over the last few decades, vocabulary has received increased attention as a key aspect of second language (L2) learning. Vocabulary knowledge is critical simply because no verbal communication is possible without words. As Read (2000, p. 1) puts it, "words are the basic building blocks, the unit of meaning from which larger structures such as sentences, paragraphs and whole texts are formed." The recognition of the present centrality of vocabulary in the field of L2 acquisition has aroused researchers' and teachers' interest in assessing vocabulary knowledge so that they can track the development of their learners' vocabulary knowledge.

Although vocabulary knowledge has been defined differently by different researchers (see, for example, Aitchison, 1994; Laufer, 1997; McCarthy, 1990; Miller, 1999; Nation, 1990, 2001; Richards, 1976), it is agreed that knowing a word involves more than knowing the relationship between its form and meaning. In order to measure vocabulary knowledge, a number of vocabulary tests have been created and validated (Beglar, 2010; Beglar & Hunt, 1999; Ishii & Schmitt, 2009; Laufer & Nation, 1999; Meara & Buxton, 1987; Nation, 1983, 1990; Nation & Beglar, 2007; Read, 1993, 1998; Schmitt, Ng, & Garras, 2011; Schmitt, Schmitt, & Clapham, 2001). These tests are of theoretical value in investigating how different aspects of vocabulary knowledge are interrelated and how vocabulary knowledge is related to other language skills such as reading and listening. They are also of practical value in providing learners with useful information on their current level of vocabulary knowledge and clearly indicating how

many words are needed for achieving a particular goal. However, existing vocabulary tests do not aim at indicating how learners can become proficient in vocabulary learning. This thesis is one of the first attempts to create such tests; that is, it aims to investigate the important prerequisites for vocabulary learning proficiency (VLP), and to develop and validate tests measuring VLP for learners of English as an L2.

1.1 What is Vocabulary Learning Proficiency?

Vocabulary learning proficiency (VLP) refers to the ability necessary to facilitate L2 vocabulary learning. It determines how efficiently words are learned and predicts learners' rate of vocabulary development. For example, as will be discussed later, affix knowledge is considered to be part of VLP, because knowing many affixes may facilitate vocabulary learning. The meanings of affixed words may easily be inferred and remembered if learners know the affix and its base. For example, if learners know the affix *un-* and the base *happy*, it should be easier for them to learn the word *unhappy* than those who do not know the affix *un-*.

The notion of VLP may be related to the broader notion of language aptitude which refers to "basic abilities that are essential to facilitate foreign language learning" (Carroll & Sapon, 1959, p. 14). The importance of language aptitude is supported by Ehrman and Oxford (1995) who showed that language aptitude as measured by the Modern Language Aptitude Test (MLAT; Carroll & Sapon, 1959) was correlated most strongly to L2 proficiency of all the individual-difference variables examined, including learning strategies, leaning styles, personality, motivation, and anxiety. Similar to the notion of language aptitude, VLP may be taken as the ability necessary to facilitate L2 vocabulary learning. It should be noted that VLP is different from language aptitude in

that it specifically deals with vocabulary learning rather than general language proficiency. This makes it possible to provide learners with diagnostic feedback on their weaknesses in vocabulary learning in particular.

VLP can be taken as one of many factors that affect the difficulty of vocabulary learning. A large number of attempts have been made to investigate factors affecting vocabulary learning in order to determine the most effective ways of learning vocabulary. These factors may be classified into four categories: textual (nature of the text in which the word is used), word (features of the word), learner (individual learner differences in knowledge, effort, strategies, etc.), and situational factors (mental tasks learners do with the word) (Paribakht & Wesche, 1999). Here are several examples of the four types of factors affecting vocabulary learning.

1. Textual factors.

- *Repetition*. The more often a word is met, the more likely it is to be learned (Horst, Cobb, & Meara, 1998; Jenkins, Matlock, & Slocum, 1989; Rott, 1999; Saragi, Nation, & Meister, 1978; Waring & Takaki, 2003; Webb, 2007a). Although the results are inconclusive as to how many encounters are needed for acquisition to occur, researchers agree that meeting a word repeatedly contributes to learning.
- *Coverage*. As learners increase their vocabulary, they have greater vocabulary coverage of unsimplified text. The lower density of unknown words makes these words more noticeable because there are only a few unknown words among a large number of known words. Greater coverage also provides learners with richer contexts to draw on when they guess the meanings of unknown words. A minimum of 95% of the words in a text may need to be known for successful

guessing to occur (Laufer, 1989; Liu & Nation, 1985), and 98% coverage may be ideal for more successful guessing (Hirsh & Nation, 1992; Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006).

- *Usefulness*. Words that are useful to understanding a text may be more likely to be learned than those that are not (Brown, 1993). This may be the reason why content words (e.g., nouns, verbs, and adjectives) tend to be learned more easily than function words (e.g., articles and prepositions) (Brown, 1993; Paribakht & Wesche, 1997).
- *Quality of context*: In order for learners to be able to correctly guess the meanings of unknown words, sufficient contextual clues need to be present in the text (e.g., Dubin & Olshtain, 1993; Haastrup, 1985; Haynes, 1993; Hulstijn, 1992; Sternberg, 1987). Without such clues, it may be difficult for successful guessing to occur, and thus the unknown words are unlikely to be learned incidentally.
- *Quantity of input*: The more input learners get, the more they meet the vocabulary. A large amount of input is necessary because vocabulary learning from meaning-focused input is a gradual process where one meeting with a word adds to the small amounts of vocabulary knowledge gained from previous meetings (Day, Omura, & Hiramatsu, 1991; Nagy, Herman, & Anderson, 1985; Pitts, White, & Krashen, 1989; Saragi, et al., 1978). This could be differentiated from repetition because a large quantity of input does not necessarily mean a large number of repetitions of lower-frequency words.

2. Word factors.

- *Sound-letter correspondence*. Word forms with clear relationships between

sounds and letters may be easy to learn. Research (Abbott, 2000) indicates that the rules of English sound-letter relationships may be acceptably reliable (the rules apply to more than 75% of the words investigated), but there are a number of words that do not follow the rules. For example, for the majority of one-syllable words with the spelling of vowel-consonant-*e*, the final *e* is silent and the vowel says its name (e.g., *cake* and *joke*), but there are some exceptions (e.g., *have* and *come*).

- *Similarity of word forms*. Words that share similar sounds or spellings (e.g., *adapt/adopt* and *industrial/industrious*) may be confusing and difficult to differentiate between. Similarity of word forms, or synformy, is a difficulty-inducing factor for learners of English regardless of their first language (L1) (Laufer, 1988, 1991).
- *Morphological transparency*. A word that consists of semantically transparent word parts may be easy to learn if learners know each of the word parts (Bauer & Nation, 1993; Nagy & Anderson, 1984). For example, the word *unhappy* may be easy to learn because it consists of semantically transparent parts (*un-* and *happy*). The word *prefix*, on the other hand, may not be as easy to learn as *unhappy* because its meaning is difficult to infer from *pre-* and *fix*. Research (Bensoussan & Laufer, 1984; Laufer & Bensoussan, 1982) indicates that L2 learners tend to misunderstand the meanings of deceptively transparent words which look as if they were composed of meaningful word parts (e.g., *outline* for 'out of line' and *discourse* for 'without direction').
- *Multiple meanings*. It may be difficult to learn all the meanings of a word with multiple meanings, because learners may not pay attention to other meanings of

the word if they already know one of its meanings. Research (Bensoussan & Laufer, 1984) showed that when guessing the meaning of an unknown word in context learners who already knew one of the meanings of a polyseme did not think of another meaning even if this meaning was not consistent with the context.

3. Learner factors.

- *Cumulative gains in vocabulary knowledge*. Vocabulary knowledge accumulates for each aspect of knowledge as vocabulary size increases. Cumulative gains in knowledge reduce the amount of learning required to learn unknown words. For example, as vocabulary size increases, learners are more likely to have known synonyms of unfamiliar words. Knowledge of those synonyms may facilitate the learning of unfamiliar ones for some aspects of vocabulary knowledge such as grammatical functions and syntagmatic associations (Webb, 2007b). Knowledge of word parts may also increase as vocabulary knowledge develops. In the initial stages of vocabulary development, learners have no knowledge of word parts. Gradually as knowledge of word parts accumulates it becomes easier to learn words which are made up of word parts.
- *Strategies*. Previous studies have identified a number of vocabulary learning strategies such as guessing from context, dictionary use, and word-pair learning (Gu & Johnson, 1996; Oxford & Crookall, 1990; Schmitt, 1997; Williams, 1985). Research generally indicates that more successful vocabulary learners tend to rely on a wider variety of strategies (Ahmed, 1989; Gu & Johnson, 1996; Lawson & Hogben, 1996; Moir & Nation, 2002).
- *L1 knowledge*. Establishing the form-meaning relationship of an L2 word may

be easier if a learner's L1 has roughly the same word form with roughly the same meaning as the L2 word. Some languages have a number of loan words and cognates shared with English. In Japanese, for example, English loan words account for 45.5% of the 3,000 most frequent words of Nation's (2004) BNC word lists, which may facilitate Japanese students' learning of English high-frequency words (Daulton, 2004).

- *Motivation*. Words may be learned more effectively when learners have stronger motivation to learn them. Research (Crookes & Schmidt, 1991; Gardner & MacIntyre, 1991) showed that a significantly greater number of words were learned when monetary rewards were given to those who scored higher than the pre-determined level. Laufer and Hulstijn (2001) subsume need under motivation for the purpose of operationalising the effectiveness of vocabulary learning tasks. Learners are more likely to learn words which they feel a need to learn. For example, beginners may not feel a need to learn synonyms. However, advanced learners may feel it is necessary to learn synonyms because they may need to express the same word in different ways.

4. Situational factors.

- *Noticing*. Learners need to notice that words are unknown. Factors that may affect noticing include the importance of the word in the context, the importance of the word to the learner (need), repetition, and L2 proficiency (Ellis, 1990; Nation, 2001; Schmidt & Frota, 1986).
- *Strength of effort*. A stronger effort to understand a text may lead to a greater depth of processing which may lead to better retention of vocabulary (Hulstijn, 1992; Joe, 1995; Laufer & Hulstijn, 2001). This is because a learner with a

strong effort tends to use multiple sources of information (e.g., context, sentence-level grammar, and background knowledge) when guessing the meanings of unknown words and checking the guesses for accuracy (de Bot, Paribakht, & Wesche, 1997; Haastrup, 1985; Nassaji, 2003).

- *Depth of processing*. The more deeply a piece of new information is processed, the more likely it is to be learned. It is argued that the depth with which the information is processed is more important to long-term memory than the length of time that the information is held in short-term memory (Craik & Lockhart, 1972; Craik & Tulving, 1975). In an attempt to operationalise the notion of depth of processing, Laufer and Hulstijn (Laufer & Hulstijn, 2001) proposed an *Involvement Load Hypothesis* which examines the effectiveness of vocabulary learning tasks. Subsequent studies (Hulstijn & Laufer, 2001; Kim, 2011) generally support this hypothesis.

VLP is related to learner factors because different learners are assumed to have different levels of VLP. Among several factors relating to learners, the present research focuses on cumulative gains in vocabulary knowledge and vocabulary learning strategies because they are teachable. Learners' existing knowledge and strategies are different from other learner factors such as L1 knowledge and motivation which affect vocabulary learning but cannot be taught. They are also different from textual, word, and situational factors in this respect. Since VLP is teachable, the results of the present research will be easily applicable to teaching in normal classroom settings. An in-depth discussion of what is involved in learners' existing knowledge and strategies will be discussed in the subsequent chapter.

VLP may be well explained in relation to Laufer and Husltijn's (2001) *Involvement Load Hypothesis*, one of the most influential theories on L2 vocabulary learning. This hypothesis predicts the relative efficacy of vocabulary learning tasks on the assumption that retention of words is conditional upon the degree of involvement in processing the words. Involvement load is quantified by totalling the ratings of the three components: need (motivation to learn words), search (attempt to find the meaning or form of a word), and evaluation (attempt to choose an appropriate form or meaning of a word by comparison with other words or other meanings of the word). Each of the three components is rated as 0 (absence of the component), 1 (presence of the component in its moderate version), or 2 (presence of the component in its strong version). It is assumed that a task with a higher involvement load (the total of the ratings from the three components) will be more effective for retention of words than that with a lower involvement load. Laufer and Hulstijn report that the Involvement Load Hypothesis is generally consistent with previous studies that examined the effects of different tasks on vocabulary learning. Supportive evidence for this hypothesis is provided by subsequent research (Hulstijn & Laufer, 2001; Kim, 2011). While involvement load is an important factor in vocabulary learning, the level of involvement required for acquisition might also be determined by learners' proficiency level of vocabulary learning. In other words, learners with a higher VLP may require lower involvement for retention of words. For example, a learner with knowledge of the affix *fore-* and the word *warn* may require lower involvement for learning the word *forewarn* than a learner without this knowledge, because this knowledge decreases the amount of knowledge required to learn *forewarn* (the pronunciation, the spelling, and the meaning of *fore-* and *warn*). In this sense, involvement load and VLP may be taken to be complementary to each other.

Thus, more effective vocabulary learning may result from both a task with a higher involvement load and a learner with a higher VLP.

The notion of VLP is also related to learning burden which was first introduced by Nation (1990, 2001). Learning burden is the amount of effort needed to learn and remember a word. If a word follows the patterns that learners are already familiar with, then the learning of the word becomes easy and the learning burden of it is light. For example, if a learner knows words such as *make* and *take*, then the learning burden of the word *lake* is light because these words share a similar pattern of pronunciation. VLP and learning burden are similar in assuming that learners' existing knowledge makes vocabulary learning easier, but are different in that focus is put on learners for VLP (how efficiently the learner can remember words) and on words for learning burden (how much effort is needed to learn the word).

This section has explained the notion of VLP by comparing it with relevant notions such as language aptitude, involvement load, and learning burden. VLP refers to the ability necessary to facilitate vocabulary learning. VLP and language aptitude are similar in this respect, but are different in that VLP focuses on vocabulary learning in particular instead of general language proficiency. Among several factors relating to learners, the present research focuses on learners' existing knowledge and strategies which are different from factors in other categories (textual, word, and situational factors) and other learner factors such as L1 knowledge and motivation which do affect vocabulary learning but cannot be taught. VLP is also related to the Involvement Load Hypothesis in that a learner with a higher VLP may require lower involvement for retention of words. Finally, VLP is related to learning burden in assuming that learners' existing knowledge makes vocabulary learning easier, but the difference between VLP

and learning burden lies in whether focus is placed on learners (VLP) or words (learning burden).

1.2 Why Is It Important to Measure VLP?

The development and validation of VLP tests is of great value, because, to my knowledge, there are no tests that aim to measure how efficiently words can be learned. Existing vocabulary tests aim to measure learners' knowledge of vocabulary, with focus being placed either on how many words are known (e.g., the Vocabulary Levels Test; Nation, 1983, 1990) or how well a word is known (e.g., the Word Associates Test; Read, 1993, 1998). These tests, however, do not tell us how learners can improve their ability to learn vocabulary. Language aptitude tests such as the MLAT include items that relate to vocabulary, but their purpose is to measure learners' aptitude for general language learning and not for vocabulary learning. This makes it difficult to provide learners with diagnostic information on what is needed to become efficient in vocabulary learning. A dearth of tests measuring VLP indicates a need for new approaches to vocabulary assessment. These tests may provide learners with diagnostic information on how to improve their VLP.

VLP tests will benefit teachers because they may diagnose their learners' vocabulary learning weaknesses. The diagnosis will provide learners with information on which types of knowledge and strategies specifically need to be learned in order to become more proficient in vocabulary learning. For example, if a VLP test indicates that a learner needs to know more about word parts, he could then direct his effort to gaining knowledge of word parts. Since teachers have little time to teach low-frequency words in class, it is important to help learners become proficient in vocabulary learning

strategies so that they can effectively continue with vocabulary learning on their own.

VLP tests may also help to determine a critical threshold after which vocabulary learning becomes significantly easier. An investigation into the relationship between learners' performance on VLP tests and their vocabulary size may indicate a general tendency that a learner with a particular vocabulary size has a particular level of VLP. For example, a learner who knows 3,000 words or more might know the affix *fore-* because it is found in *forecast* and *foresee* which are within the 3,000-word level in the British National Corpus (BNC) word lists (Nation, 2004). Knowledge of *fore-* might in turn facilitate the learning of less frequent words such as *forewarn* and *forego*. If the goal of vocabulary learning were set at developing a vocabulary size of 8,000 words, which might be necessary to achieve the 98% coverage of written text (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006), then the VLP level of learners with a vocabulary size of 8,000 words might be taken as the threshold levels for efficient vocabulary learning.

VLP tests may contribute to a better understanding of L2 vocabulary learning. Previous studies have investigated the relationships between learners' existing knowledge/strategies and vocabulary learning; for example, existing phonological knowledge relates to vocabulary learning (e.g., Hulme, Maughan, & Brown, 1991) and existing word part knowledge does too (e.g., Schmitt & Meara, 1997). However, few attempts have been made to collectively approach the notion of VLP; that is, previous research has focused on only specific areas of learning proficiency and remains to be synthesized from a theoretical and practical perspective. This thesis attempts to contribute to the theory of L2 vocabulary acquisition by providing validated measures of VLP and allowing empirical research into VLP.

1.3 Purpose and Scope of the Present Research

This thesis aims to develop diagnostic tests of VLP. More specifically, it creates and validates two tests of VLP: guessing from context and word part knowledge. (What components are included in VLP and why these two are chosen are discussed in Chapter 2.) In this thesis, this issue is investigated by tackling the following research questions:

1. How is the skill of guessing from context measured?
2. Does the test of guessing from context produce valid and reliable results?
3. How is knowledge of word parts measured?
4. Does the test of word part knowledge produce valida and reliable results?

In order to effectively approach these questions, this thesis consists of seven chapters. This introductory chapter is followed by a literature review (Chapter 2) which provides an in-depth discussion about what is involved in VLP. Chapter 3 explains the rationale and the procedure for creating a guessing-from-context test, one of two components of VLP tests. Chapter 4 describes ways in which poorly written items are identified and dealt with, and presents evidence for the validity of the guessing-from-context test. It also proposes new forms of the test that would be useful for both research and practical purposes. Chapter 5 provides the rationale and the procedure for creating a word part test, the second component of VLP tests. Chapter 6 describes two studies one of which was conducted to identify and rewrite poorly written items on the word part test. The other study was carried out to examine whether the written items work well for learners with a wide variety of L1 backgrounds. This chapter also proposes new forms of the word part test that would be useful to researchers and teachers. Chapter 7 provides

concluding remarks including a general discussion of the research, limitations and suggestions for future research.

The value of the present research lies in the introduction of the notion of VLP and the creation of two validated tests of VLP (guessing from context and word part knowledge). Since the present research is one of the first attempts to create measures of VLP, it focuses on the two most important components of VLP (The reasons for the choice of the two components are discussed in Chapter 2). A complete set of VLP tests will include many other components such as phonological knowledge and dictionary use, but the development of the tests of these components is beyond the scope of this research and is left for future studies.

This chapter has discussed the purpose and the importance of the present research. The subsequent chapter takes a closer look at what kinds of knowledge and strategies are involved in VLP.

CHAPTER 2

LITERATURE REVIEW

This chapter first reviews the L2 literature on what is involved in knowing a word in order to clarify which aspects of vocabulary knowledge become easier to learn with a higher VLP. It then discusses what is involved in VLP and why this thesis focuses on the skill of guessing from context and knowledge of word parts.

2.1. Vocabulary Knowledge

This section summarises aspects of vocabulary knowledge that are proposed by previous research in the field of L2 vocabulary acquisition, and discusses which aspect of vocabulary knowledge is focused on in this thesis. Table 1 presents previous studies on what is involved in knowing a word.

Table 1. Summary of what is involved in knowing a word

Richards (1976)	Nation (1990)	Laufer (1997)	Nation (2001)
Form and its derivations	Spoken form	Form (spoken)	Spoken form
	Written form	Form (written)	Written form
		Word structure	Word parts
Semantic value	Concept	Meaning	Form and meaning
Multiple meanings			Concept and referents
Associations	Associations	Lexical relations	Associations
Syntactic behaviour	Grammatical patterns	Syntactic pattern	Grammatical functions
	Collocations	Common collocations	Collocations
Frequency	Appropriateness		Constraints on use
Limitations on use	Frequency		

As shown in Table 1, all of these studies have pointed out that knowing a word involves multiple aspects of word knowledge. Important prerequisites for VLP might be different according to which aspect of vocabulary knowledge is being learned. For example, phonological knowledge might facilitate the learning of the pronunciation of a word, but it might hardly contribute to the learning of the grammatical function of a word.

Among various aspects of vocabulary knowledge, the present research focuses on the form-meaning relationship because this aspect is arguably the most important. First, the importance of the form-meaning relationship may be seen in the fact that this aspect of vocabulary knowledge is included in all previous studies with different labels. It is termed as "semantic value" by Richards (1976), "concept" by Nation (1990), "meaning" by Laufer (1997), and "form and meaning" by Nation (2001). All of these terms refer to the relationship between word form and its meaning.

Second, words are primarily units of meaning and knowledge of form-meaning relationships may be more important than other aspects of vocabulary knowledge such as grammatical function and associations because semantic knowledge is required for comprehension. Laufer, et al. (2004) argue for the centrality of the form-meaning relationship as follows:

> [A] student who knows what 'evidence' means, but does not know that it is used as a singular noun and says *'The judge listened to many evidences' will be understood, in spite of the grammatical error. On the other hand, a student who knows that 'evidence' is used in the singular but confuses its meaning with 'avoidance' will experience a breakdown in communication. (p.205)

In terms of communication where meaning is conveyed between the speaker and the listener, a grammatically incorrect sentence consisting of words with correct meanings may be more acceptable than a grammatically correct sentence consisting of words with incorrect meanings; hence, knowledge of form-meaning relationships is of particular

importance.

Third, the majority of learning materials, activities, and vocabulary tests have focused on knowledge of form-meaning relationships, perhaps because the first step in vocabulary learning is seen as establishing initial form-meaning relationships (Schmitt, 2008). For example, researchers have created and validated tests of vocabulary size (how many words are known) which are designed to measure the amount of knowledge of form-meaning relationships. The Vocabulary Levels Test (Beglar & Hunt, 1999; Nation, 1983, 1990; Schmitt, et al., 2001) requires learners to match a word meaning to its form. Here is an example.

1. business
2. clock __part of a house
3. horse __animal with four legs
4. pencil __something used for writing
5. shoe
6. wall

On this test, learners choose the correct word form that goes with each of the three meanings from a set of six options. This format directly measures the form-meaning relationship. Another test of vocabulary size is the Vocabulary Size Test (Beglar, 2010; Nation & Beglar, 2007) which requires learners to match a word form to its meaning. Here is an example.

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

On this test, learners choose the correct meaning of the target word (*miniature*) from a set of four options. This format also directly measures knowledge of form-meaning relationships. Meara and Buxton (1987) proposed a yes/no format instead of a multiple-choice format for measuring vocabulary size. They presented learners with a list of real

and nonsense words and asked them to tick the words that they knew the meaning of. If they ticked nonsense words, their scores were downgraded. The yes/no format may also measure knowledge of form-meaning relationships because learners are asked to examine whether they know the meanings of the words and the results showed that this format was significantly correlated to a multiple-choice format (r=.703, p<.001) where learners matched a word form to its meaning .

It could be argued that other existing vocabulary tests such as the Productive Vocabulary Levels Test and the Word Associates Test also measure knowledge of form-meaning relationships. The Productive Vocabulary Levels Test (Laufer & Nation, 1999) is a test in which learners have to write a word starting with a few pre-determined letters in a sentence. Here is an example.

The garden was full of fra___ flowers.

In this example, learners write the word that starts with *fra* and best fits the context. This test may be related to knowledge of form-meaning relationships because learners first determine the meaning of the blank from the context and then recall the form linked to the meaning. Another existing vocabulary test is the Vocabulary Knowledge Scale (Wesche & Paribakht 1996) which asks learners to evaluate a list of words by choosing their level of knowledge from the following five options: (1) “I don’t remember having seen this word before,” (2) “I have seen this word before, but I don’t know what it means,” (3) “I have seen this word before, and I think it means (synonym or translation),” (4) “I know this word. It means (synonym or translation),” and (5) “I can use this word in a sentence (write a sentence).” This scale is also related to the form-meaning relationship because it measures how well learners know the meanings of the words. The Word Associates Test (Read, 1993, 1995; Schmitt, et al., 2011) is a test

of knowledge of word associations. Here is an example of this test.

fundamental

neutral core perfect root	marriage objective agreement news

In this example, the target word is *fundamental* and learners choose two associates from each of the two boxes (the left box contains paradigmatic associations and the right box contains syntagmatic associations). This test is designed as a measure of depth of vocabulary knowledge, but may also be related to knowledge of form-meaning relationships, because learners are less likely to be able to choose correct answers without knowledge of the meaning of *fundamental*. To sum up, existing vocabulary tests, either explicitly or implicitly, measure knowledge of form-meaning relationships, perhaps because the form-meaning relationship is central to vocabulary knowledge.

Finally, vocabulary size, or the number of words whose meanings are known, plays a critical role in language skills such as reading and listening. Research (Nation, 2006) indicates that a vocabulary size of 8,000-9,000 words may be necessary for understanding written texts, and a vocabulary size of 5,000-6,000 words are needed for aural comprehension. Learners with a vocabulary size below these levels might have trouble understanding written or spoken texts. Research shows that vocabulary size has a strong relationship with reading comprehension. Laufer (1992) reported that positive correlations were found between reading comprehension as measured by two standardised reading tests (the reading component of *Examen Hoger Algemeen Vortgezet Onderwijs* consisting of two texts and 20 multiple-choice comprehension questions, and the English subtest of the Israeli university psychometric entrance test consisting of texts with 40 multiple-choice comprehension questions) and vocabulary

size as measured by the Vocabulary Levels Test (Nation, 1983) (r=.50, p<.01) and as measured by the Eurocentres Vocabulary Test (Meara & Jones, 1990) (r=.75, p<.01). Qian (1999) examined the relationships among reading comprehension, vocabulary size, association knowledge, and morphological knowledge, showing that reading comprehension as measured by a reading section of TOEFL (Educational Testing Service, 1987, pp. 93-100) positively correlated to vocabulary size as measured by the Vocabulary Levels Test (Nation, 1983, 1990) (r=.78, p<.05). He also found that the correlation between reading comprehension and vocabulary size was roughly as high as that between reading comprehension and association knowledge as measured by the Word Associates Format (Read, 1993, 1995) (r=.82, p<.05) and was higher than that between reading comprehension and morphological knowledge as measured by a self-made test where learners were asked to define ten words with particular affixes and then to write the part of speech of these ten words (r=.64, p<.05). These studies indicate the relative importance of knowledge of form-meaning relationships in language skills such as reading and listening.

For the above-mentioned reasons, the form-meaning relationship is arguably the most important aspect of vocabulary knowledge. The present research, thus, focuses on the learning of form-meaning relationships when referring to VLP. The subsequent section discusses what is involved in VLP; that is, what kinds of knowledge and strategies contribute to the efficient learning of form-meaning relationships.

2.2 What Is Involved in VLP?

In order to establish the form-meaning relationship of a word, learners need to gain knowledge of the word form and its meaning. The learning of unknown word forms

would be facilitated if learners already knew rules for spoken and written forms. For example, absolute beginners without any knowledge of English may have difficulty learning the pronunciation of the word *date*, but learners with knowledge of words such as *take*, *make*, and *name* may be able to learn the pronunciation of *date* easily.

The learning of word meaning may be easier if learners are more successful in deriving the appropriate meaning when they come across unknown words while reading or listening. The strategies for deriving word meaning include word part analysis, guessing from context, and consulting a dictionary (de Bot, et al., 1997; Fraser, 1999; Mori & Nagy, 1999; Paribakht & Wesche, 1999). The ability of intentional learning through L2-L1 word pairs may also be of great importance, because adult learners already possess a well-established L1 conceptual and lexical system, and L1 use may reduce the learning burden of L2 meaning at the initial stage of vocabulary development (e.g., Jiang, 2004).

Taken together, VLP involves the following six aspects:

(1) knowledge of a sound system,

(2) knowledge of sound-spelling relationships,

(3) knowledge of word parts,

(4) guessing from context,

(5) dictionary use, and

(6) word-pair learning.

The following subsections review the literature on the effects of each of the six aspects of VLP on vocabulary learning and on the effectiveness of teaching each aspect.

2.2.1. Knowledge of a Sound System

The importance of phonological knowledge in vocabulary learning is strongly supported by research on the relationship between phonological short- and long-term memory. It has been pointed out that short-term memory as measured by non-word repetition (accuracy of repeating unfamiliar spoken words) and articulatory suppression (interruption by repetition of a nonsense word during learning) affects the learning of novel foreign words and nonsense words (Ellis & Sinclair, 1996; Gathercole & Baddeley, 1989; Masoura & Gathercole, 1999; Papagno, Valentine, & Baddeley, 1991; Service, 1992). As learners' vocabulary grows, phonological long-tem knowledge becomes more important in vocabulary learning than short-term memory. Gathercole (1995) observed that for any given length of nonsense word, English-like words (e.g., *defermication*) were easier for L1 English children than non-English-like words (e.g., *perplisteronk*). Further analysis showed that short-term memory (as measured by tests of digit span and one- and three-syllable span) was more closely related to non-word repetition accuracy for the non-English-like than for the English-like words. These findings indicate that while totally unfamiliar words are largely dependent on phonological short-term memory, the learning of English-like items is likely to be facilitated by long-term lexical knowledge. Cheung (1996), in a study with Hong Kong 7th graders learning English, found that phonological short-term memory as measured by non-word repetition was related to vocabulary acquisition only for those with a small English vocabulary size. Masoura and Gathercole (2005) found that Greek children's speed of learning English words in a paired-associate learning task was strongly affected by their current English vocabulary knowledge, arguing that learners with considerable familiarity with the L2 benefit from the use of existing knowledge

representations. Gathercole, Service, Hitch, Adams, and Martin (1999) argue that long-term memory has an impact on short-term memory in the way that it helps to reconstruct words from incomplete phonological memory traces at the point of retrieval by constraining possible sequences of sounds with reference to phonotactic regularity.

Phonological knowledge as a facilitating factor seems to be segmentalised. Research (Fowler, 1991; Metsala & Walley, 1998; Walley, 1993) indicates that as vocabulary grows, children's phonological representations become increasingly segmentalised and eventually phoneme-level representations arise. The segmental nature of existing phonological representations would in turn facilitate the learning of phonological form (Bowey, 1996, 2001; de Jong, Seveke, & van Veen, 2000; Snowling, Goulandris, Bowlby, & Howell, 1986). Ellis (2001) argues that "phonology [...] develop[s] hierarchically by repeated cycles of differentiation and integration of chunks of sequences" (p.41).

The ability to segment speech sounds is called *phonological sensitivity* (or *phonological awareness*), and research shows that phonological sensitivity is improved by training. Lundberg, Frost and Petersen (1988) showed that Danish preschool children improved their phonological awareness after metalinguistic games and exercises that stimulated them to discover and attend to phonological structures. The positive effect of training on the improvement of phonological awareness is also confirmed by Byrne and Fielding- Barnsley (1995) and de Jong et al. (2000).

2.2.2. Knowledge of Sound-Spelling Relationships

In English, spelling and pronunciation are closely related to each other, and it would be of value to deal with them together. The English language uses phonograms, and

spelling ability has found to be most strongly affected by learners' phonological representation (e.g., Bradley & Huxford, 1994). On the other hand, it has been shown that spelling knowledge improves learners' memory for pronunciation (e.g., Rosenthal & Ehri, 2008).

Although English seems to have complex relationships between the sounds and the spellings that they represent, there are rules for the sound-spelling relationships in English. Abbott (2000) showed that the rules of English sound-letter relationships were reliable (the reliability was 75% or more), indicating the effectiveness of phonic knowledge in vocabulary learning.

The effectiveness of teaching phonics has been confirmed by studies with children. Bruck, Treiman, Caravolas, Genesee, and Cassar (1998) found that children with phonics instruction produced more accurate word spellings than children without phonics instruction when asked to learn and spell a list of words; in addition, the phonics children produced more conventional and phonologically acceptable patterns for the spellings of nonsense words. Similar results were obtained by Roberts and Meiring (2006). Nation (2009) argues that while most learning of L2 sound-spelling relationships occurs incidentally, deliberate teaching would help speed up the learning.

2.2.3. Knowledge of Word Parts

The usefulness of word parts has been underlined by corpus-based research. Nagy and Anderson (1984) analysed a 7,260-word sample from the Word Frequency Book (Carroll, Davies, & Richman, 1971), and found that each base form has on average between 1.57 and 3.22 derived forms (excluding inflected forms) depending on the way in which a word is counted as a family. Of course, word parts are not necessarily equal

in value. Research into the frequency of affixes has shown that only a small number of affixes appear frequently (e.g., Thorndike, 1941).

The importance of morphological knowledge is also supported by research from a psychological perspective; that is, the relationship between word stems and their morphologically related forms is psychologically real. Nagy, Anderson, Schommer, Scott, and Stallman (1989) conducted a lexical decision task with 95 L1 English speakers in the United States and showed that the speed with which a word was recognised was conditional upon the total frequency of its morphologically related forms rather than the frequency of the word itself. This indicates that morphologically related words are linked to one another in the mental lexicon and that every word does not have a completely separate entry.

Empirical studies with L2 learners of English have shown that knowledge of word parts positively correlates with vocabulary size. Schmitt and Meara (1997), in a study with 95 Japanese university students learning English, reported a moderate correlation between affix knowledge and vocabulary size (r = .27-.41). Higher correlations are reported in subsequent studies such as Qian (1999) (r = .69), Mochizuki and Aizawa (2000) (r = .54-.65), and Ishii and Schmitt (2009) (r = .73).

The importance of explicit instruction of word parts has been pointed out by Bauer and Nation (1993) and Nation (1990, 2001). This is empirically supported by Schmitt and Zimmerman (2002) who indicated that learners might not acquire word part knowledge automatically through exposure. They argue that word parts need to be learned explicitly, especially for productive use.

2.2.4. Guessing from Context

The skill of guessing the meanings of unknown words from context plays an important part in learning vocabulary through reading and listening. Research in foreign language acquisition (Brown, Waring, & Donkaewbua, 2008; Day, et al., 1991; Dupuy & Krashen, 1993; Horst, et al., 1998; Hulstijn, 1992; Pitts, et al., 1989; Waring & Takaki, 2003) indicates that words may be successfully inferred from context, but only a small proportion of words may be retained. Guessing from context has the advantage of providing learners with the meaning of a word in particular use. Given that many words are polysemous and the meaning of a word is largely determined by the context in which it occurs (Miller, 1999), guessing from context may be an effective way of gaining knowledge of meaning (Anderson & Nagy, 1991).

Research (Fukkink & de Glopper, 1998; Kuhn & Stahl, 1998; Walters, 2004) indicates that instruction has a positive effect on the guessing strategy, and that the effectiveness of instruction may vary according to proficiency level. Walters (2006) found that less proficient learners benefited most from general strategy instruction (presenting a general rule for guessing followed by practice), while more advanced learners benefited most from context instruction (making learners aware of specific types of context clues).

2.2.5 Dictionary Use

Research (Chun & Plass, 1996; Hill & Laufer, 2003; Hulstijn, Hollander, & Greidanus, 1996; Knight, 1994; Laufer & Hill, 2000; Luppescu & Day, 1993; Peters, 2007) has indicated that the use of dictionaries contributes to gains in vocabulary knowledge. Luppescu and Day (1993), for example, examined the effects of using bilingual

dictionaries while reading on vocabulary learning with 293 Japanese university students learning English as a foreign language. The results showed that those who used a dictionary scored significantly higher on a subsequent vocabulary test than those who did not. The results also indicated that some learners were unable to locate the appropriate meaning of a word that was looked up in a dictionary. Knight (1994) investigated the effects of dictionary use on vocabulary learning with 112 intermediate Spanish learners of English, showing that those who used a dictionary scored higher on both an immediate and delayed (two weeks later) vocabulary tests than those who did not. Similar results were obtained by Hulstijn et al. (1996) with 78 Dutch advanced students of French, although those who were given marginal glosses (L1 translations of unknown words) scored higher on a subsequent vocabulary test than those who were allowed to use a bilingual dictionary.

Although dictionary use may have a positive effect on vocabulary learning, learners may not be efficient at it. Tono (1988) examined the skill of dictionary use by Japanese university students with a low to intermediate level of proficiency by measuring multiple aspects of dictionary use including pronunciation, spelling, part of speech, meaning, reference speed, derivatives, synonyms, usage, and social background. The results showed that the participants were successful in deriving the appropriate meaning of 67-71% of the words that were looked up in a dictionary. The results also showed that the participants performed better for some aspects of dictionary use (e.g., success rate of finding inflected forms = 78%) than others (e.g., success rate of finding derivatives = 46%). Fraser (1999) examined eight Francophone university students' strategies for dealing with unknown words while reading, and found that the participants were successful in deriving the appropriate meaning of 78% of the words

that were looked up in a dictionary. These studies indicate that there is still room for improving learners' skill of dictionary use.

The skill of dictionary use may be improved by instruction. Fraser (1999) reported that the participants became slightly more successful in deriving the appropriate meanings of unknown words in a dictionary after metacognitive strategy training (raising awareness of the use of lexical processing strategies such as consulting a dictionary, guessing from context, and ignoring). More systematic strategy training may focus on each aspect of what is involved in dictionary use such as Schofield's (1982) seven steps in using a dictionary for comprehension.

2.2.6. Word-Pair Learning

Although deliberate, decontextualised word-pair learning has often been considered to be a less useful activity than contextualised learning (e.g., Oxford & Crookall, 1990), it is of great importance because it enables learners to focus on particular words that meet their needs and to control how often the words are encountered so that they may be effectively stored in memory (Nation, 2001). L2 empirical research shows that deliberate learning leads to greater and faster gains of form-meaning relationships than incidental learning does. Prince (1996) found that learning with L1 translations was more effective than contextualised learning in the number of newly learned words recalled. Laufer and Shmueli (1997) showed that words presented in a list were learned better than words presented in context. These studies indicate that deliberate learning should be seen as complementary to incidental learning, rather than as an inferior method of learning.

The value of deliberate learning is also supported by a recent study (Elgort, 2007)

which showed that implicit knowledge as measured by primed lexical decision tasks resulted from intentional decontextualised learning. This indicates that deliberate learning of vocabulary may lead to the kind of knowledge required for normal language use.

Word-pair learning ability is likely to be improved through instruction. Reviewing the literature, Nation (2001) proposes some teachable guidelines for effective deliberate learning. For example, the guideline *use recall* is based on research findings (e.g., Baddeley, 1990) indicating that retrieving the meaning of an unknown word is more effective than seeing the word and its meaning at the same time.

Sections 2.2.1-2.2.6 have looked at each of the six aspects of VLP: knowledge of a sound system, knowledge of sound-spelling relationships, knowledge of word parts, guessing from context, dictionary use, and word-pair learning. Previous studies have generally indicated that these six types of knowledge and strategies have a positive effect on vocabulary learning, and are improved by teaching. Taken together, these six aspects of VLP may be taken as important prerequisites for efficient vocabulary learning. Among the six aspects of VLP, the present research focuses on guessing from context and knowledge of word parts. The subsequent section discusses the reasons for that.

2.3 Importance of Guessing from Context and Knowledge of Word Parts

Since it is impossible to deal with the creation and validation of the tests of all six aspects of VLP in a single PhD, this thesis focuses on two of them: guessing from context and knowledge of word parts. This section discusses the reasons why it might be more important to measure the skill of guessing from context and knowledge of word parts than other aspects of VLP.

The first reason for the creation of the two tests is that there have been no validated tests that measure the skill of guessing from context and knowledge of word parts. Existing tests of guessing from context may not be useful for detecting learners' weaknesses, because the tests are not easy to complete and grade (e.g., Haastrup, 1991) or the tests do not measure multiple aspects of guessing (e.g., Schatz & Baldwin, 1986) (see Section 3.4 for a review of the existing tests of guessing from context). Existing tests of word part knowledge may also have limitations, because the tests do not measure multiple aspects of word part knowledge (e.g., Wysocki & Jenkins, 1987) or a limited number of word parts are measured (e.g., Schmitt & Meara, 1997) (see Sections 5.5.2.1., 5.5.3.1, and 5.5.4.1 for a review of the existing tests of word part knowledge).

The skill of guessing from context plays a key role in vocabulary learning, because it is the most frequent and preferred strategy when learners deal with unknown words in context. Paribakht and Wesche (1999) conducted an introspective study with ten intermediate ESL learners with various L1 backgrounds, and found that the participants used guessing from context for 80% of the unknown words whose meanings they actively tried to identify. Fraser (1999) also conducted an introspective study with eight Francophone university students, and found that guessing was the most frequent strategy (44%) of all the strategies employed (consult = 29%; ignore = 24%; other = 3%). Cooper (1999) examined strategy use for dealing with unknown idioms with 18 ESL learners with a variety of L1 backgrounds, and found that guessing was the most frequent strategy (28%) of all the strategies employed.

Although guessing is the most frequent strategy for dealing with unknown words in context, learners' guesses often result in failure. Nassaji (2003), in a study with 21 ESL learners with various L1 backgrounds, found that they made correct guesses for

25.6% of all unknown words and 44.2% of these items if partially correct guesses were included. Similar results were obtained in a study by Parry (1991), in which the four participants' ratios of correct guesses to the total number of guesses ranged from 12% to 33%. These low success rates suggest that there is much room for improvement in the guessing strategy. Creating a guessing-from-context test may diagnose learners' weaknesses in guessing and contribute to the improvement of the guessing strategy.

Guessing from context may be a major source of vocabulary learning where vocabulary learning occurs while reading and listening. The importance of vocabulary learning through reading and listening can be seen in a number of previous studies that provide positive but modest evidence for gains in vocabulary knowledge for both L1 acquisition (Jenkins, Stein, & Wysocki, 1984; Nagy, Anderson, & Herman, 1987; Nagy, et al., 1985; Shu, Anderson, & Zhang, 1995) and L2 acquisition (Brown, et al., 2008; Day, et al., 1991; Dupuy & Krashen, 1993; Horst, et al., 1998; Hulstijn, 1992; Pitts, et al., 1989; Waring & Takaki, 2003). It is argued that the vast majority of words are learned while reading and listening especially for L1 acquisition (Nagy & Anderson, 1984). For L2 acquisition, vocabulary learning through reading and listening may become important especially when learners have gained knowledge of high-frequency words that appear in a wide variety of texts. This is because teachers may have little time to deal with a daunting number of low-frequency words in class and learners may need to increase their vocabulary knowledge on their own while reading or listening to the texts that are of interest to them. Taken together, the improved skill of guessing has the potential to facilitate vocabulary learning through reading or listening, because learners rely on the guessing strategy most frequently when dealing with unknown words in context and good guessers are likely to have a greater opportunity to derive the

appropriate meaning of an unknown word and learn it.

The second aspect of VLP examined in this thesis is knowledge of word parts. For English learners, this knowledge is of great value because about half of English words are morphologically complex (Anglin, 1993; Goulden, Nation, & Read, 1990; Nagy & Anderson, 1984). In addition, a corpus-based study by Nagy and Anderson (1984) indicated that an increasingly greater proportion of semantically transparent words appear as the word frequency goes down. This suggests that knowledge of word parts is useful especially for the learning of low-frequency words which may not be taught in class and thus need to be learned independently.

Another advantage of word part knowledge is that it may help learners check whether an unknown word has been successfully guessed from context (Mori, 2002; Mori & Nagy, 1999; Nagy & Anderson, 1984; Nation, 2001). In other words, integration of information from context and word parts may make guessing more successful and contribute to effective vocabulary learning. Being aware of the two sources of information is of great importance because word meanings may not easily be determined by a single source of information. Research (Beck, McKeown, & McCaslin, 1983; Schatz & Baldwin, 1986) indicates that contextual clues are not always sufficient for deriving the meanings of unknown words. Research (Bensoussan & Laufer, 1984) also indicates that learners may be misled by some deceptive word parts (e.g., *bother* is not *both* + *-er*).

Finally, knowledge of word parts and the skill of guessing from context may be the most useful strategies for learners because these strategies may be used in any situation for any words. Both strategies do not require supplementary materials such as word cards and flash card software. The skill of guessing is available when learners read or

listen to virtually any material (e.g., novels, newspapers, and movies) and thus learning occurs both inside and outside the classroom. It may also make learning pleasurable because learners can read or listen to any material that is interesting to them. Knowledge of word parts is also useful because it is widely available for learning word families (morphologically and semantically related words). Research (Nagy & Anderson, 1984) shows that every word has more than one derivative on average (1.57 – 3.22 derivatives excluding inflections depending on the difference in the definition of semantic relatedness), indicating that knowledge of affixes may help expand vocabulary rapidly.

For the above-mentioned reasons, this thesis created and validated tests of guessing from context and knowledge of word parts. An in-depth review of the literature on how these two aspects of VLP have been measured will be provided separately in Chapter 3 for guessing from context and Chapter 5 for knowledge of word parts so that it may be easier to recognise the difference between the formats used in previous studies and the format used in this thesis.

2.4 Summary

This chapter has made the purpose and the scope of this thesis clearer by justifying the need to measure two aspects of VLP: the skill of guessing from context and knowledge of word parts. It first argued that the form-meaning relationship should be examined among the various aspects of vocabulary knowledge because this aspect is arguably the most important. Narrowing an argument down to one aspect of vocabulary knowledge was necessary because important prerequisites for VLP may be different according to the aspect of vocabulary knowledge being learned. A review of the literature indicates

that the VLP for form-meaning relationships involves the following six types: knowledge of a sound system, knowledge of sound-spelling relationships, knowledge of word parts, guessing from context, dictionary use, and word-pair learning. All these six types of VLP facilitate vocabulary learning and are improved by teaching. Among the six types of VLP, this thesis focuses on guessing from context and knowledge of word parts. The reasons for the choice of measuring the skill of guessing from context are (1) that there have been no established tests of guessing from context; (2) that guessing is the most frequent strategy when learners come across unknown words in context; (3) that L2 learners often fail to derive the appropriate meaning of an unknown word from context; (4) that guessing may be the main source of independent vocabulary learning; and (5) that guessing is one of the most useful strategies for learners. The reasons for the choice of measuring knowledge of word parts are (1) that there have been no established tests of word part knowledge; (2) that knowledge of word parts may facilitate the learning of affixed words which account for about half of the words in English; (3) that combining information from word parts and context may make guessing more accurate; and (4) that word part knowledge is one of the most useful strategies for learnes. The subsequent chapter reviews the literature on how guessing from context has been measured and looks at the procedure for creating a guessing-from-context test.

CHAPTER 3

DEVELOPMENT OF THE GUESSING FROM CONTEXT TEST

This chapter aims to describe the procedure for developing a guessing-from-context test. After discussing the scope of the test, it focuses on particular types of clues that can be used in guessing from context. It then provides an in-depth discussion on the process for determining the format of the test.

3.1 Scope of the Research

The present guessing-from-context test (GCT) aims to measure how well L2 learners can guess the meaning of unknown words in written text. Some studies (e.g., Carton, 1971; Nassaji, 2003) prefer the term *inferencing* to *guessing*, because the term *guessing* may be taken to imply random guesswork based on arbitrary intuition. In the present research, the term *guessing* refers to informed guessing which does not include the notion of such random guesswork. The term *inferencing* is avoided to differentiate between the act of deriving meaning from context which often results in failure and the act of drawing a conclusion from formal reasoning based on available data which is often used as a technical word in the field of science and logic (e.g., *statistical inference* in science and *valid inference* in logic).

It is important to distinguish between deriving the meaning of an unknown word from context and learning it because successful guessing does not always lead to learning (e.g., Fraser, 1999). The present research focuses on guessing from context instead of learning. Guessing from context plays a critical role in vocabulary learning

because it is an important prerequisite for vocabulary learning while reading and is the main strategy used when learners meet unknown words in context (de Bot, et al., 1997; Fraser, 1999; Hulstijn, 1992; Paribakht & Wesche, 1999). It should be reasonable to assume that more efficient learners in guessing from context have a greater chance to learn words while reading.

The GCT aims to provide learners with diagnostic information on their weaknesses in guessing from context. In so doing, it measures whether they can find and use clues in context. Among various types of clues, it deals with grammar (part of speech of the unknown word) and discourse (relationships with other words or phrases in the context) clues. The subsequent section reviews what types of clues are available to L2 learners and discusses the reasons for focusing on grammar and discourse clues.

3.2 Clues for Guessing from Context

This section reviews what types of clues have been found to be available to L2 learners when they guess the meaning of unknown words from context. It also discusses why the GCT measures knowledge of grammar and discourse.

Carton (1971) logically proposed three categories for cues[1] that can be used in guessing from context: intra-lingual, inter-lingual, and extra-lingual. Intra-lingual cues come from knowledge of the target language, including morphological, syntactic, and phonological knowledge. Inter-lingual cues are based on knowledge of languages other than L2 (L1 and others) including loan words and cognates. Finally, extra-lingual (or contextual) cues include knowledge of the world.

Borrowing the tripartite taxonomy of cue types from Carton (1971), Haastrup

[1] Carton preferred the term *cue* to *clue*. The two terms are used interchangeably in this thesis.

(1985, 1987, 1991) empirically investigated how L2 learners guessed from context based on introspective and retrospective data from Danish learners of English. The results indicated that each of the three cue types could be subdivided into two as shown in Table 2. The taxonomy is not mutually exclusive.

Table 2. Taxonomy of cue types by Haastrup (1985, 1987, 1991)

Cue type	Sub-category	Description
Contextual	1. The co-text	One or two words from the immediate co-text; the immediate co-text; a specific part of the co-text beyond the sentence of the test word; unspecified use of the co-text.
	2. Knowledge of the world	Factual knowledge; attitudes; beliefs; prejudices.
Intralingual	1. The test word	Phonology/orthography; morphology; lexis; word class; collocations; semantics.
	2. The syntax of the sentence	Structure of the sentence in which the test word occurs.
Interlingual	1. The L1 (Danish)	Phonology; orthography; morphology; collocations; semantics.
	2. Ln (other than L1 and L2)	General reflections; morphology; lexis; semantics.

Using the think-aloud method with 10 ESL university students, de Bot, et al. (1997) identified eight knowledge sources used in guessing which varied widely in frequency of use. Table 3 presents the eight types of knowledge in descending order of frequency (how often each knowledge source was used).

Table 3. Taxonomy of knowledge sources by de Bot, et al. (1997)

Knowledge source	Description	Frequency (%)
Sentence-level grammatical knowledge	Parts of speech in a sentence.	34.6
Word morphology	English derivations and inflections.	15.0
Punctuation	Punctuation and capitalisation rules.	11.2
World knowledge	Knowledge of the theme and topic.	9.3
Discourse and text	Information from other parts of the text.	3.7
Homonymy	Phonetic similarities between the target word and another word (e.g., *melt* and *smell*).	3.7
Word associations	Words associated with the target word (e.g., *accommodation* and *hotel reservation*).	2.8
Cognates	Word cognates, mainly between English and French in their study.	1.9
Unknown	-	17.8

As shown in Table 3, some knowledge sources were used more frequently than others. The most popular two sources were sentence-level grammar and morphology, accounting for half of the sources used for guessing.

Based on introspective and retrospective data from 21 intermediate ESL learners, Nassaji (2003) identified five knowledge sources. Table 4 presents these knowledge sources in descending order of frequency (how often each knowledge source was used). It also provides the percentage of successful guesses including partial success for each knowledge source.

Table 4. Taxonomy of knowledge sources by Nassaji (2003)

Knowledge source	Description	Frequency (%)	Success (%)*
World	Knowledge of the content or the topic that goes beyond what is in the text.	46.2	54.2
Morphological	Knowledge of word formation and word structure, including word derivations, inflections, word stems, suffixes, and prefixes.	26.9	57.1
Grammatical	Knowledge of grammatical functions or syntactic categories such as verbs, adjectives, or adverbs.	11.5	41.7
Discourse	Knowledge about the relation between or within sentences and the devices that make connections between the different parts of the text.	8.7	55.6
L1	Knowledge of similar words in the L1.	6.7	42.9

*Percentage of successful guesses

As shown in Table 4, some knowledge sources were used more frequently than others, which supports the findings of de Bot, et al. (1997). It should be noted that the frequency of world knowledge was widely different: While de Bot, et al. found that their participants relied heavily on grammatical (34.6%) and morphological knowledge (15.0%) followed by world knowledge (9.3%), Nassaji's participants used world knowledge most frequently (46.2%) followed by morphological (26.9%) and grammatical knowledge (11.5%). This may have been due to the nature of context, supporting Nation's (2001, p.257) argument that background clues are not always present. Table 4 also shows that the percentages of success ranged between 41.7% and 57.1%, suggesting that the probability of success in guessing might vary according to the knowledge source used.

Previous studies (de Bot, et al., 1997; Haastrup, 1985, 1987, 1991; Nassaji, 2003)

largely agree on the types of clues that L2 learners use in guessing. Table 5 summarises the clues identified by these studies with the aim of showing the overlap among them. The overlapping categories are listed in the same row; for example, "The co-text" used by Haastrup, "Discourse and text" and "Punctuation" used by de Bot, et al., and "Discourse knowledge" used by Nassaji refer to similar notions.

Table 5. Summary of clue types

	Haastrup (1985, 1987, 1991)	de Bot, et al. (1997)	Nassaji (2003)
Type 1	The co-text	Discourse and text Punctuation	Discourse knowledge
Type 2	Knowledge of the world	World knowledge	World knowledge
Type 3	The test word	Word morphology Word associations Homonymy	Morphological knowledge
Type 4	The syntax of the sentence	Sentence-level grammatical knowledge	Grammatical knowledge
Type 5	The L1 (Danish)	Cognates	L1 knowledge
Type 6	L3, L4, etc.		

As shown in Table 5, clues for guessing may be categorised into six types. Among those clue types, discourse (Type 1) and grammar (Type 4) clues were selected for the GCT based on the following two criteria:

1. The clue can be taught; and
2. The clue can be used in every context.

The first criterion was set up so that teachers could help learners improve their skill of guessing from context based on the GCT. L1 knowledge (Type 5) and knowledge of another language (Type 6) did not meet this criterion because language teachers are not

always familiar with their students' L1s and third languages (L3s).

The second criterion presupposes that clues that are always present in context are more useful than those that are not. World knowledge (Type 2) did not meet this criterion, because world knowledge is not always available especially when learners read about unfamiliar topics. It may also be outside the scope of language teachers because they cannot be familiar with every topic that their students may encounter or know the extent of their students' world knowledge. Another type of knowledge that did not meet the second criterion was knowledge of the test words. The use of word clues is not always available or helpful. Morphological knowledge cannot be used effectively when an unknown word does not consist of analysable word parts. It has also been pointed out that wrong guesses are typically caused by the misunderstanding of the word forms (Bensoussan & Laufer, 1984; Laufer & Sim, 1985; Nassaji, 2003). For example, Bensoussan and Laufer (1984) reported that some L2 learners wrongly guessed the meaning of *outline* as 'out of line' by breaking it into parts. As Nation (2001, p. 259) suggests, it may be more effective to use word form clues as a supportive aid for checking a guess rather than as a main strategy for guessing.

The GCT focuses on discourse (Type 1) and grammar (Type 4) clues. A discourse clue is a clue found in other parts of the context. A grammar clue refers to the part of speech of the unknown word which makes it possible to analyse the structure of the sentence in which the unknown word is used. There are three reasons for measuring knowledge of these two types of clues on the GCT. First, research has shown that the skills of using discourse clues (e.g., Fukkink & de Glopper, 1998; Kuhn & Stahl, 1998; Walters, 2006) and analysing the grammatical structure in a sentence (e.g., Carpay, 1974; van Parreren, 1975) can be improved by teaching. These two types of knowledge

are different from other types of knowledge such as L1 and L3 which are difficult to teach.

Second, although grammar and discourse clues may not always be helpful (Beck, et al., 1983; Schatz & Baldwin, 1986), they are present in every context: An unknown word always has a grammatical function in a sentence and is used in discourse. These clues are different from other clues such as morphological and world knowledge which are not always present.

Finally, the studies on procedures for guessing from context essentially underscore the importance of grammar and discourse. For example, Clarke and Nation (1980) proposed a five-step procedure for guessing from context which was later expanded by Nation and Coady (1988) and Nation (1990, 2001). Here are the five steps.

Step 1: Decide on the part of speech of the unknown word.
Step 2: Look at the immediate context (the sentence in which the word is used).
Step 3: Look at the wider context (the relationship with other sentences).
Step 4: Guess.
Step 5: Check the guess.

In Clarke and Nation's procedure, Step 1 focuses on grammar clues and Steps 2 and 3 involve discourse clues. The use of background knowledge is not included in their procedure because it is not always available and is less likely to lead to vocabulary learning. The use of word part knowledge is included in Step 5 for checking the guess because word part analysis is sometimes misleading.

A similar procedure was proposed by Williams (1985) who classified the guessing strategy into the following four categories:

1. Work out the unfamiliar word's part of speech.
2. Search the context for other words that will help you to puzzle out the meaning of the new word.
3. Let those other words throw light on the meaning of the new word.
4. Try your inference, to check that it makes sense.

In Williams' procedure, the first strategy focuses on indentifying the part of speech of unknown words. The second strategy deals with discourse clues. Similar to Clarke and Nation's (1980) procedure, guessing and checking the meaning come last.

For the purpose of guiding teachers to help students guess the meaning of unknown words in classroom activities, Bruton and Samuda (1981) proposed six stages for guessing the meaning of unknown words.

Stage 1: Focusing on the word to guess.
Stage 2: Getting students to guess the meaning of the word.
Stage 3: Asking about clues available in the passage.
Stage 4: Justifying the acceptable guesses made by students.
Stage 5: Providing the precise meaning of the word.
Stage 6: Providing back-up exercises.

Bruton and Samuda's guessing strategy is different from other strategies proposed by Clarke and Nation (1980) and Williams (1985) in that guessing comes earlier than finding information for guessing. Despite this difference, Bruton and Samuda suggested that in Stage 3 teachers should make their students find grammar and discourse clues in the passage. For example, in order to guess the meaning of the unknown word *neglected* in the sentence *In the morning, Carter found a letter in the most obvious place of all, which he had somehow neglected*, learners need to recognise the relative pronoun (*which* refers to *place*) and the tense (*neglected* occurred before *found*).

This section has reviewed the empirical studies that provided a taxonomy of clues that L2 learners use in guessing from context. Despite different labels for clue types, previouse studies generally agreed on the categorisation of clues. The GCT focused on grammar and discourse clues because they are teachable and usable in every context. These two types of clues are also included in previous studies that proposed a procedure for guessing from context. The subsequent section addressees what is involved in grammar and discourse clues.

3.3 Clues in Context

This section provides an in-depth discussion of what is involved in grammar and discourse clues and how these clues contribute to deriving the meanings of unknown words.

3.3.1 Grammar

Knowledge of grammar helps learners identify the part of speech of an unknown word. Clarke and Nation (1980, p. 212) argue that knowing the part of speech is important because it allows the "Who does what to whom?" analysis. For example, in the sentence *Typhoon Vera killed or injured 218 people and crippled the seaport city of Keelung* (*crippled* is the target word to be guessed), the unknown word *crippled* is a verb. Based on the grammatical analysis of the sentence, learners may recognise that Typhoon Vera did something (=crippled) to Keelung. What a typhoon does is likely to have a negative influence on a city. This analysis may not be sufficient to arrive at the precise meaning of *cripple*, but together with the phrase *killed or injured 218 people* which is connected with the coordinate conjunction *and*, learners may be able to guess its meaning as 'damage' or 'destroy'. Clarke and Nation also emphasise the importance of grammar by arguing that failures in guessing seem to be frequently caused by misunderstanding the part of speech of the unknown word. For example, although *laterally* is an adverb, a learner may guess its meaning as 'coming after or later' which is an adjective rather than an adverb. Recognising the part of speech of *laterally* may help learners avoid this mistake.

The GCT controlled for the parts of speech of the test words (words to be guessed) because part of speech might affect the chance of success in guessing from context.

Aborn, Rubenstein, and Sterling (1959) showed that the percentage of words guessed successfully varied according to part of speech, and suggested the difficulty order as follows: adjectives, nouns, adverbs, and verbs with adjectives being the most difficult. Dulin (1969) reported on a difficulty hierarchy in the order of verbs, adjectives, adverbs, and nouns with verbs being the most difficult. Liu and Nation (1985) found a different difficulty order: adjectives, adverbs, nouns, and verbs with adjectives being the most difficult. These studies do not agree as to the difficulty order of parts of speech perhaps because many factors other than part of speech were involved such as the density of unknown words and the types of clues available in context. However, their findings indicate that success in guessing may be affected by the part of speech of the test word.

The GCT focuses on nouns, verbs, adjectives, and adverbs, because these four parts of speech account for the vast majority of word types in English. The four classes of words contrast with function words such as prepositions and auxiliary verbs which are small in number, have little lexical meaning, and indicate the mood of the speaker or the grammatical relationship with other words.

The number of test words for each part of speech was based on frequency data in the BNC so that the proportion would reflect authentic language. The frequency ratio of (noun): (verb): (adjective): (adverb) was 9:6:3:2 in the BNC (Leech, Rayson, & Wilson, 2001). This ratio was used for selecting test words for the GCT.

3.3.2 Discourse

The importance of discourse clues has been underscored in studies on L2 learners' guessing process (de Bot, et al., 1997; Haastrup, 1985, 1987, 1991; Nassaji, 2003) and practical models for guessing from context (Bruton & Samuda, 1981; Clarke & Nation,

1980). Despite the importance of using discourse clues for successful guessing, research has revealed that even L1 high-school, undergraduate and graduate students are not aware of a variety of discourse clues (McCullough, 1943; Strang, 1944). Research has also indicated that success in the use of discourse clues may depend on age and the type of clues (e.g., explicitness and closeness) (Ames, 1970; Carnine, Kameenui, & Coyle, 1984; Dulin, 1969; Quealy, 1969; Rankin & Overholser, 1969).

A number of attempts have been made to classify discourse clues by analysing a) various texts (Artley, 1943; Deighton, 1959; Dulin, 1970; Johnson & Pearson, 1984; Spache & Berg, 1955; Walters, 2006), b) data from learners who guessed the meanings of real words that they reported as unknown (McCullough, 1943, 1945, 1958), and c) data from learners who guessed the meanings of nonsense words or blanks (Ames, 1966; Seibert, 1945). The taxonomies of discourse clues proposed by the previous studies vary widely. Some clues (e.g., direct description) are included in all the studies, while others (e.g., example) are not. Some clues with different labels refer to largely the same notion (e.g., direct explanation and definition).

In order to measure L2 learners' overall guessing ability, the GCT included a wide variety of discourse clues. This was because some discourse clues might be easier to use than others. For example, Carnine, et al. (1984) found that more explicit clues (e.g., synonyms) were easier to use than less explicit clues (e.g., indirect descriptions). Table 6 summarises the discourse clues identified by nine studies (Ames, 1966; Artley, 1943; Deighton, 1959; Dulin, 1970; Johnson & Pearson, 1984; McCullough, 1945; Seibert, 1945; Spache & Berg, 1955; Walters, 2006) in descending order of frequency (how many studies mentioned the clue). Similar categories are listed in the same row.

As shown in Table 6, discourse clues may largely be classified into twelve types.[2] The clues are not mutually exclusive.

Table 6. Summary of discourse clues

	Artley (1943)	McCullough (1945)	Seibert (1945)	Spache & Berg (1955)	Deighton (1959)
direct description	direct explanation	definition	definition or description	direct explanation	definition
indirect description	inference subjective clues figures of speech	mood or situation	clues found in the general meaning of the paragraph	inference tone or mood figures of speech	inference
contrast/ comparison	antonym	comparison and contrast	antonym opposite ideas comparison	structural	
synonym	synonym	synonym	synonym idea repeated in two forms		
appositive	appositive interpolation			structural	
modification	non-restrictive form			structural	modifiers
restatement			sentence which follows		restatement
cause/effect		summary			
words in series			series a chain of actions		
reference					
association			frequently coupled		
example					example

[2] Another type of taxonomy was proposed by Sternberg and Powell (1983) who classified context clues based on the type of information that the context conveys, rather than the type of devices used to convey the information. Their taxonomy involved the following cues: temporal (duration, frequency or time), spatial (location), value (worth or desirability), stative descriptive (physical property such as size, shape, colour, odour, feel, etc.), functional descriptive (purpose, action or use), causal/enablement (cause or enabling conditions), class membership, and equivalence (synonymy or antonymy) cues.

Table 6. (cont'd)

	Ames (1966)	Dulin (1970)	Johnson & Pearson (1978)	Walters (2006)
direct description	definition and description question-and-answer pattern	direct description	direct definitions or explanations	description
indirect description	tone, setting or mood main idea and supporting details preposition	tone or mood	inferences subjective clues figures of speech	inference
contrast/ comparison	comparison and contrast	contrast	comparisons or contrasts	contrast
synonym	synonym	linked synonyms and appositives	substitute words	
appositive	non-restrictive and appositive	linked synonyms and appositives		punctuation
modification	modification			adjective (phrase, clause)
restatement			restatement	restatement in another clause restatement in the same clause
cause/effect	cause and effect	cause-effect relationships	summary	
words in series	words connected in series			grouping
reference	reference			reference
association	association			
example				examples and illustrations

Table 6 focuses on discourse clues (clues found in other parts of the context), and thus excludes other types of clues such as morphology (Seibert, 1945), world knowledge (Ames, 1966; Artley, 1943; Johnson & Pearson, 1984; McCullough, 1945; Seibert, 1945), familiar expressions (Ames, 1966; Artley, 1943; Johnson & Pearson, 1984; McCullough, 1945), and typography (e.g., italics and quotation marks) (Artley, 1943).

Here are some brief explanations about each discourse clue (Underlined words are the test words to be guessed. Some studies used real words, while others used blanks or

nonsense words for the test words.).

1. *Direct description*. All nine studies mentioned this type of clue using a variety of labels such as (direct) description, explanation, and definition. In direct description, the unknown word is explicitly defined by words such as *mean* and *is*. Here are some examples.

    a) Many objects are buoyant which simply means that they will float on the surface of the water. (Artley, 1943, p. 71)
    b) Material or physical things are of course things that we can touch, see, taste or feel. (Deighton, 1959, p. 6)

2. *Indirect description*. This type of clue does not provide any explicit signal words for guessing. Learners need to guess based on the information around it. Here are some examples.

    a) Tom was a foot taller and thirty pounds heavier than Kirk. He overwhelmed him in the match. (Johnson & Pearson, 1984, p. 117)
    b) A little later, as he sped northward along a California cliotol, Kendricks was stopped by a highway patrol officer. (Ames, 1966, p. 78)

    This category also includes figures of speech and subjective clues such as mood, tone or setting. Figure of speech includes using a word or words with a different meaning from the usual meaning, being divided into simile which always includes the words *like* or *as* and metaphor which does not. The subjective clues are based on tone or mood in the context. For example, the author may want to express a character in a context as happy, angry or sad. Below are the examples of simile (c), metaphor (d), and subjective clues (e).

    c) The old car lurched forward like an anxious dog released by its master. (Artley, 1943, p. 70)
    d) The trail snaked its way through the hills, winding sinuously from one pass to another. (Spache & Berg, 1955, p. 111)
    e) I was alone. The day was dull with black clouds overhead. The dreary landscape cast a spell of melancholy over me. (Johnson & Pearson, 1984, p. 117)

3. *Contrast/comparison*. The meaning of the unknown word is typically the opposite of that of a familiar word, phrase, or idea that is contrasted or compared with it. Antonyms are included in this category because the meaning of the unknown word is the opposite of the antonym. This type of clue is often marked with words or phrases such as *in contrast, rather than, instead of, unlike, but* and *or*. Here are some examples.

    a) Rather than his usual mood of cheerful good humor, today his manner appeared quite dour. (Dulin, 1970, p. 442)
    b) The argument became more than just a simple disagreement but progressed rapidly into a distasteful affray. (Spache & Berg, 1955, p. 110)

4. *Synonym*. This type of clue is a known synonym for the unknown word. The synonym clues are often marked with words such as *too* and *also*. In other cases, synonyms are used in similar sentence structures.

    a) Could it be parl to baccarat, too?[3] (Ames, 1966, p. 72)
    b) When Jim heard that his bicycle would be ready that evening, he was __. He was glad that he would have it in time for the trip with Tom the next day. (McCullough, 1945, p. 3)

5. *Appositive*. The unknown word is explained in the word or phrase following it. Appositive is typically marked with commas, colons, semicolons, and dashes. Here are some examples.

    a) The fertilizer should supply plenty of vegetable matter, which by decaying furnishes humus, the food for plant life. (Artley, 1943, p. 69)
    b) The invading armies proceeded to ravage – completely ruin and destroy – the local churches, schools, and public buildings. (Dulin, 1970, p. 442)

6. *Modification*. The unknown word is modified by a word, phrase or clause, typically by an adjective clause which is marked with relatives such as *which, who,* and *that*.

[3] Here is an example of a successful guesser's response to show the context of this sentence. "It is the idea of the end so perhaps *by-by* would be right. He has used this phrasing above in referring to the end of blackjack and the word 'too' gives me the idea of his repeating this phrasing" (Ames, 1966, p.72).

Here are some examples.

a) The decaying vegetable matter of the fertilizer will furnish humus, which is food upon which plant life depends. (Artley, 1943, p. 69)
b) One clue is given by metabolism tests which measure the rate at which the chemical and physical processes in the body are carried on and at which energy is produced and utilized. (Deighton, 1959, p. 7)

7. *Restatement*. The unknown word is restated in the preceding or following word, phrase or sentence. Restatement is often signalled with words or phrases such as *or*, *that is*, and *in other words*. In many other cases, no explicit signals are given. Here are some examples.

a) The cockroach is an insect that has two antennae, or feelers, on its head. (Johnson & Pearson, 1984, p. 117)
b) And his consecutive games record went on and on. Sick or well, he never missed a game. (Deighton, 1959, p. 15)

The restatement clue overlaps with the synonym clue to a large extent. In the present research, restatement is taken as the clue that is restated in another sentence such as the example b). Restated words or phrases such as the example a) are taken as part of the synonym clue.

8. *Cause/effect*. The unknown word in the cause may be logically guessed from the effect, or vice versa. The cause/effect relationships are marked with words such as *because*, *since*, *as*, *thus* and *therefore*. Here are some examples.

a) Since he was determined that he would finish the task no matter how long it took, he worked doggedly on. (Dulin, 1970, p. 443)
b) He reads not for fun but to improve his mind and render his conversation less caxall. (Ames, 1966, p. 80)

Summary clues labelled by McCullough (1945) and Johnson and Pearson (1984) were included in the cause/effect clues, because their examples could be taken as cause/effect relationships.

c) His knees shook and his eyes seemed to pop as he looked all around, for he was very much ___. (McCullough, 1945, p. 3)
d) Being an itinerant preacher, my grandfather travelled through all parts of the state. (Johnson & Pearson, 1984, p. 117)

9. *Words in series*. The unknown word is part of a series of words, phrases, or ideas, typically connected with the word *and*. Here are some examples.

a) Under questioning, Kendricks broke down and mespolded the policeman's murder. (Ames, 1966, p. 70)
b) Shrimp, clams, oysters and ___ are all at risk during certain months of the year from a certain bacteria in the water called 'the Red Tide'. (Walters, 2006, p. 182)

10. *Reference*. The meaning of the unknown word may be derived by unlocking referral words such as *this, that* and *it*. Here are some examples.

a) Look at the figures for deaths that occur at birth, or during the first year of life, for every 1000 infants in these countries.

Sweden 15.3
U.S. 25.3

These whafarbins carry an unpleasant message. (Ames, 1966, p. 75)
b) In 1962, in the Rocky Mountains, near Denver, Colorado, water was forced through pipes into a layer of rocks 4000 meters below the surface of the ground. Shortly after this ___ of water, there was a small number of earthquakes. (Walters, 2006, p. 182)

11. *Association*. The meaning of the unknown word may be derived by association with a word around it. The most frequent association clues are links between a noun and a verb and between an adjective and a noun. Here are some examples.

a) He heard the crack of the whip. (Seibert, 1945, p. 306)
b) "In our reader," my oldest child once snorted, "all the little boys wear short nerns and their names all end in 'y' and they're cute." (Ames, 1966, p. 76)

12. *Example*. The unknown word is explained with an example or included in an example explaining a familiar idea. The example clue is typically marked with words or phrases such as *like*, *for example*, and *such as*. Here is an example (Walters (2006) mentions an example clue, but no example is given).

> Girls on the average consistently do better in the test items involving esthetic response such as matching colors and shapes and discriminating in pictures. (Deighton, 1959, p. 6)

In an attempt to include a wide variety of discourse clues, the GCT deals with all twelve types of clues.

In summary, the GCT focuses on the four parts of speech (noun, verb, adjective, and adverb) and the twelve types of discourse clues (direct description, indirect description, contrast/comparison, synonym, appositive, modification, restatement, cause/effect, words in series, reference, association, and example). Before discussing how these aspects of guessing from context are measured in the GCT, the subsequent section looks at how the guessing skill has been measured in previous studies.

3.4 Previous Tests Measuring Guessing from Context

This section reviews how the skill of guessing from context has been measured. A number of studies have employed think-aloud protocols which require learners to verbalise what they think while guessing the meanings of unknown words from context. For the purpose of identifying types of contextual clues, Ames (1966) replaced every 50th words with nonsense words and asked the participants to guess the meanings of the nonsense words aloud. Here is an example.

> I wonder how much the security of the country is being safeguarded by the paunchy reservist who spends one evening a week at the Reserve center *thacing* the fat with the boys, thereby escaping from the dishes at home.

In this example, the italicised word *thacing* is the test word to be guessed. Each participant was asked to respond with a synonym or a definition of the target word, and then to explain what part of the text helped him or her guess its meaning. This technique was also used by Quealy (1969), Rankin and Overholser (1969), and Haynes (1993).

Other researchers (Arden-Close, 1993; Fukkink, Blok, & de Glopper, 2001; Haastrup, 1987, 1991; Huckin & Bloch, 1993; Laufer & Sim, 1985; Morrison, 1996; Nassaji, 2003; Parry, 1991) also used think-aloud protocols but used real words instead of nonsense words. One of the advantages of this format is that it is sensitive to partial gains in vocabulary knowledge. For example, Nassaji (2003) classified the participants' answers into three categories: successful, partially successful, and unsuccessful. Another advantage is that it may provide learners with diagnostic information about their weaknesses in guessing. For example, teachers may recognise that some learners do not make use of a wide variety of contextual clues which may help derive the meaning of unknown words. However, it takes too much time to administer the test because think-aloud techniques typically require researchers to observe each learner's responses individually. It is also difficult to grade the test objectively because a variety of answers may be possible (e.g., synonyms and definitions). This indicates a need for a test that is easy to administer and grade.

Another way of measuring the skill of guessing from context is to use a multiple-choice format which requires learners to choose the meaning of the target words from a set of options. Here is an example used by Schatz and Baldwin (1986). The italicised word *ruefully* is the target word to be guessed.

> He takes out an envelope from a drawer, and takes paper money from it. He looks at it *ruefully*, and then with decision puts it into his pocket, with decision takes down his hat. Then dressed, with indecision looks out of the window to the house of Mrs. Lithebe, and shakes his head.
>
> RUEFULLY
> (A) sorrowfully
> (B) thankfully
> (C) fearfully
> (D) casually
> (E) longingly

In this example, learners must choose the meaning of *ruefully* from five options. Carnine, et al. (1984) also used a multiple-choice format where learners must choose the meaning of the target word from four options. Recognising the incremental nature of vocabulary learning, other researchers (Nagy, et al., 1987; Nagy, et al., 1985) created three levels of multiple-choice items for each target word. The level of difficulty was determined based on the similarity in meaning between the target word and the distractors. The items at the most difficult level require a clear understanding of the meaning of the target words, while those at the easiest level require a vague understanding of it because the distractors were created so that they would be as dissimilar as possible even in terms of part of speech. These studies indicate that a multiple-choice format is easier to administer and grade than a think-aloud technique because the studies with the former format tended to have a greater number of participants than those with the latter technique. However, the multiple-choice formats used in the previous studies do not provide any information about how learners may improve their guessing skill because it measures only one aspect of guessing; that is, deriving the meaning of unknown words. It is unclear from this format why a learner was not successful in deriving the meaning of unknown words. In order for the GCT to be a useful tool for improving learners' VLP, this problem needs to be resolved.

In summary, the problems with previous tests of guessing from context include 1) inability to identify learners' weaknesses and 2) administrative difficulty. In the present research, the test format was determined so that the above-mentioned problems may be resolved. The format of the GCT meets the following criteria:

1. The test identifies learners' weaknesses in guessing; and
2. The test is easy to complete and grade.

In order to meet the first criterion, the GCT measures three aspects of the skill of

guessing from context: knowledge of part of speech, contextual clues, and meaning. Measuring these aspects is of practical value because the GCT may provide learners with diagnostic information about their weaknesses in guessing. In order to meet the second criterion, the GCT is written in a multiple-choice format. This allows easy administration and objective and quick grading. The subsequent sections provide an in-depth discussion on the procedure for selecting test words, creating passages, and writing items for each of the three sections.

3.5 Creation of Contexts

This section discusses how contexts were created for the GCT. More specifically, it describes the procedure for selecting test words and creating reading passages.

3.5.1 Selection of Test Words

The test words to be guessed from context were randomly selected from low-frequency words which were listed between the 11th and 14th 1,000 word families in the BNC word lists developed by Nation (2006). Low-frequency words were used to minimise the likelihood that test-takers would know the words. Known words replaced by blanks or nonsense words were not considered appropriate, because knowledge of idioms and collocations might affect guessing. For example, learners may find it easy to guess the blank in the sentence *She burst into* ____ *when she heard the sad news*, if they know the frequently used expression *burst into tears*. This may measure knowledge of idioms or collocations instead of the ability to guess from context.

Among various word types in a word family,[4] the most frequent word type in the

[4] A word family includes Levels 1 to 6 of Bauer and Nation's (1993) affix levels.

BNC was chosen as the test word. For example, the past tense form *absconded* was chosen from its family which was made up of the following members: *abscond, absconded, absconding, absconds, absconder,* and *absconders*. This is because *absconded* was the most frequent of all the word types in its family. This was to maximise the likelihood that the items would represent typically encountered unknown words from their frequency levels.

The test words were nouns, verbs, adjectives, and adverbs. As discussed earlier, the number of test words for each part of speech was based on the following ratio: (noun): (verb): (adjective): (adverb) = 9:6:3:2.

The selected low-frequency words were replaced by nonsense words to make sure that test-takers had no prior knowledge of the word forms. Each nonsense word was created by changing the consonants of a low-frequency word and had roughly the same number of letters and the same inflectional and derivational suffixes as the original word. For example, *absconded* was changed into *turmilted* which was created by changing the consonants of *burnished* which was listed in the 14th 1,000 word families in the BNC word lists. *Burnished* had the same number of letters and the same inflectional suffix *-ed* as *absconded*. The nonsense words had roughly the same number of letters as the original words in order to reflect the word length of the original words which might affect success in guessing (Laufer, 1997). The nonsense words also had the same inflectional (e.g., *-ed* and *-s*) and derivational suffixes (e.g., *-ly* and *-ness*) as the original words in order to indicate their syntactic properties. Without suffixes it may be difficult to identify the part of speech of the unknown word in a sentence. For example, it may be easy to recognise that the nonsense word *ronditly* is an adverb in the sentence *He found the book ronditly while walking*, because it ends with *-ly* and appears in the

place where an adverb can occur. Without the suffix *-ly*, it may be difficult to determine the part of speech of *rondit* in the sentence *He found the book rondit while walking*. *Rondit* may also be taken as a noun which means *shop* or *sale*. The suffixes were the ones listed in Levels 2 to 6 of Bauer and Nation's (1993) affix levels.

3.5.2 Reading Passages

In research on guessing from context, there are typically two ways of presenting test words in passages. One is to present multiple test words in a reading passage which contains several paragraphs, and the other is to present one test word in a reading passage which consists of one or more sentences. Most previous studies (e.g., Ames, 1966; Haastrup, 1991; Nassaji, 2003; Quealy, 1969) used the former way which may reflect actual reading situations where learners encounter several unknown words in reading. However, this method is not appropriate for the GCT for three reasons. First, it is difficult to select semantically unrelated test words. As the test words are chosen from the passage written about a particular topic, the test words would essentially be related to each other in meaning. If the topic deals with a conceptually difficult notion, the test words might also be conceptually difficult words which in many cases are difficult to guess (Graves, 1984; Jenkins & Dixon, 1983; Nagy, et al., 1987). In addition, the measurement of learners' ability of guessing from context might be affected by their familiarity with the topic which is not the focus of the GCT. With other things being equal, those who know the topic better may get higher scores than those who do not.

Second, it is difficult to include various types of discourse clues in a small number of longer passages. With limited types of discourse clues, it is difficult to measure learners' overall ability of guessing from context. One solution to this problem may be

to use a fixed-word cloze where every fixed number of words is deleted. For example, Ames (1966) replaced every 50th word (if it was a content word) by a nonsense word. This method ensures that the words are selected without any arbitrary intuitions and discourse clues to the nonsense words are not biased towards particular clues. However, in many cases, the test words were known words replaced by nonsense words whose meanings may be easy to guess based on knowledge of idioms or collocations.

Finally, local independence of items may be violated. Local independence, which is necessary for latent variable models such as the Rasch model, requires that "the success or failure on any item should not depend on the success or failure on any other item" (Bond & Fox, 2007, p. 172). Suppose test-takers must guess two unknown words in a passage. Those who guess one unknown word correctly may be more likely to be successful in guessing the other, if the contexts surrounding the two unknown words are related to each other.

For the reasons above, the GCT includes one unknown word per reading passage. This format makes it possible to include various types of words from various topics so that the effect of background knowledge may be minimised. Moreover, contexts that contain a wide variety of discourse clues are more effectively included. This format also guarantees local independence of items because each test word is embedded in a different passage. A potential weakness of this format is that each passage needs to be relatively short in order to include a sufficient number of items for obtaining reliable results. Short passages fail to measure whether learners can utilise global clues which are found in a remote place such as a different paragraph. However, immediate clues may be much more important than global ones, because previous research indicates that in many cases learners arrive at successful guessing based on immediate rather than

global clues and poor guessers may not be able to use information in immediate contexts to guess unknown words (Haynes, 1993; Morrison, 1996).

In the GCT, each passage consisted of 50-60 running words in order to provide sufficient words for guessing the unknown words. Research (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006) has indicated that knowledge of 98% or more words surrounding an unknown word is desirable for successful guessing to occur. Each passage had 50 or more words in order to arrive at the 98% coverage. At the same time, it had 60 or less words in order to minimise the test time per item.

Each passage was selected from a paragraph containing the test word in the BNC. It was carefully chosen to exclude passages that require special topic knowledge (e.g., specialist periodicals and journals) because the GCT does not aim to measure knowledge of topics. The passages were selected so that the twelve discourse clues mentioned earlier would be evenly distributed. The place of the discourse clues was carefully controlled because the proximity of clues to the unknown words might affect the success in guessing (Carnine, et al., 1984). In so doing, half of the discourse clues appeared in the same sentences as the test words, and the others appeared outside the sentences containing the test words.

In order to maximise the likelihood that the passages would be comprehensible to test-takers, the vocabulary used in the GCT was controlled. The passages were simplified so that they consisted of words from the most frequent 1,000 word families in the BNC word lists.[5] Simplification was made to get rid of low-frequency words, and not to change the content or discourse clues.

In summary, each passage 1) had one test word, 2) was selected from the BNC, 3)

[5] Some passages contained words that were not included in the most frequent 1,000 word families (e.g., *smell* and *wine*) due to the difficulty of paraphrasing these words. A pilot study was conducted to see whether learners had trouble in understanding the passage due to these words (see Section 3.7).

consisted of 50-60 words, 4) had one discourse clue either within or outside the sentence containing the test word, and 5) was simplified so that all the running words were from the most frequent 1,000 word families. Here is an example of how a reading passage was created. The test word is *connoisseur* which is in the 11th 1,000 word families in the BNC word lists. Below is the original passage taken from the BNC (The test word *connoisseur* is underlined).

> However, the most powerful response of all to the food is to its smell, or fragrance. This is the really important information cats are receiving when they approach a meal. It is why many will sniff it and then walk away without even attempting to taste it. Like a wine connoisseur who only has to sniff the vintage to know how good it is, a cat can learn all it wants to know without actually trying the food. (Source: "Catlore". Morris, Desmond. London: Cape, 1989)

This passage includes a modification clue: the test word is modified by the relative clause that follows it. This original passage was changed by 1) replacing the test word *connoisseur* with the nonsense word *candintock*, 2) replacing lower-frequency words with higher-frequency words (preferably the most frequent 1,000 word families), and 3) limiting the passage to 50-60 running words. Here is the modified passage.

> Cats have a good nose for food. Many cats smell food and then walk away without even trying it. Like a wine candintock who only has to smell the wine to know how good it is, a cat can learn all it wants to know without actually eating the food.

The first two sentences in the original passage *However, the most [...] approach a meal* were simplified into the first sentence in the modified passage *Cats have a good nose for food*, in order to limit the context to 60 words. In the following sentence, *sniff* was changed to *smell*, and *attempting to taste* to *trying*. In the last sentence, *sniff the vintage* was changed to *smell the wine*.[6]

[6] All the words in the passage are listed in the most frequent 1,000 word families in the BNC word lists except for two words (*smell* and *wine*) which are listed in the second most frequent 1,000 word families.

This section looked at how contexts were created for the GCT. The subsequent section discusses the test format used for the GCT.

3.6 Test Format

This section addresses the test format common to all the question types and then discusses the format specific to each question type.

3.6.1 General Format

The aim of the GCT is to measure the skills of deriving the meanings of unknown words and using grammar and discourse clues for guessing. To meet this purpose, each passage was followed by the following three questions: part of speech (identifying the part of speech of an unknown word), contextual clue (finding the contextual clue that helps guess its meaning), and meaning (deriving the unknown word's meaning).

The test items were written using a multiple-choice format (choosing an answer from a set of options) instead of a recall format (writing an answer), because 1) it is easily completed and graded; 2) it is sensitive to partial knowledge (recognition tends to be easier than recall) (Laufer, et al., 2004; Laufer & Goldstein, 2004); 3) it is familiar to learners with various L1 backgrounds; and 4) poorly written items can be identified based on item analysis.

No *Don't know* options were provided because their use might depend on test-takers' personality. Some people may prefer to choose answers even for difficult items by elimination or random guessing, whereas others may prefer to stop thinking about difficult ones and choose *Don't know*. The scoring of *Don't know* responses is also difficult. If *Don't know* is regarded as a wrong answer, risk-takers gain more benefit

than non-risk-takers. It is also not appropriate to regard *Don't know* responses as missing data because the responses are not missing but reveal something about test-takers' knowledge.

All of the part of speech questions had a fixed set of four options (noun, verb, adjective, and adverb), because the part of speech of each test word was one of these four options. The clue and the meaning questions, on the other hand, had three options. Rodriguez (2005) argues that three options are optimal for the multiple-choice format based on his meta-analysis of 80 years of research. The main reason for the preference of three options is that a greater number of three-option items can be administered per unit of time than four-option items, leading to improvement on test validity. Reducing the number of options from four to three has little effect on item difficulty and test reliability. Although test-takers have a 33% chance of getting a correct answer by random guessing, previous studies (Costin, 1970, 1972; Kolstad, Briggs, & Kolstad, 1985) indicate that such random guessing rarely occurs and the quality of the distractors rather than the number of distractors plays a crucial role in the effective suppression of random guessing.

The order of the three questions was arranged as follows: part of speech, contextual clue, and meaning. This was based on Clarke and Nation's (1980) procedure for guessing: determine the part of speech of the unknown word, look for clues in context, and guess. Guessing came last in order to minimise a learning effect from earlier to subsequent questions. For example, suppose the meaning question came first. If the options of a meaning question were all verb meanings such as 'to do something', test-takers might find it easy to choose *verb* for the part of speech question without examining the sentence structure. If the options of a meaning question were related to a

particular phrase or sentence in the passage, they might easily choose that phrase or sentence for the clue question.

3.6.2 Part of Speech

The part of speech question aims to measure whether test-takers can recognise the part of speech of the test word. Every item had the following four options: noun, verb, adjective, and adverb. In the example below, test-takers are asked to choose the part of speech of the test word *candintock* (original word: *connoisseur*) from four options. The test word is written in bold and underlined so that the test-takers can recognise it with ease.

> Cats have a good nose for food. Many cats smell food and then walk away without even trying it. Like a wine **candintock** who only has to smell the wine to know how good it is, a cat can learn all it wants to know without actually eating the food.
>
> (1) noun (2) verb (3) adjective (4) adverb

The correct answer is Option 1 where the test word is the complement of the preposition *like*.

Here is another example. The test word is *decontanically* (original word: *orthographically*). This nonsense word has the typical adverbial suffix *-ly* in order to help identify the part of speech.

> When we try to look at the process of reading carefully, we will meet a further problem. Some words sound like other words, even though they are **decontanically** different. An example would be the words "see" and "sea." These two words sound exactly the same, but they include different letters.
>
> (1) noun (2) verb (3) adjective (4) adverb

The correct answer is Option 4 because the test word has the *-ly* ending and occurs between the verb *are* and the adjective *different*.

3.6.3 Contextual Clue

The contextual clue question aimed to measure whether test-takers can find a discourse clue which helps them to guess the meaning of the unknown word. Each item had three options: one correct answer and two distractors. The correct answer was the phrase or sentence that included one of the twelve discourse clues selected for the GCT. The distractors were taken from the phrases or sentences that were not helpful in guessing the meaning of the test word. In order to control the effect of proximity to the test word (Carnine, et al., 1984), one distractor was chosen from the sentence containing the test word, and the other was chosen from outside the sentence. If the sentence containing the test word was too short to create a distractor, the two distractors were chosen from outside the sentence. The distractors were also created so that the length of the distractors was roughly the same as that of the correct answer in order to make sure that the length would not indicate a correct answer. Here is an example of a clue question. Test-takers are asked to choose the word or phrase that can help them to work out the meaning of the test word from the three options underlined in the passage.

> Cats have a good nose for food. Many cats smell food and then (1)walk away without even trying it. Like a wine **candintock** (2)who only has to smell the wine to know how good it is, (3)a cat can learn all it wants to know without actually eating the food.
>
> (1) walk away without even trying it
> (2) who only has to smell the wine to know how good it is
> (3) a cat can learn all it wants to know without actually eating the food

The correct answer is Option 2 where the relative clause modifies the test word. The two distractors (Options 1 and 3) are roughly the same in length as the correct answer. One distractor (Option 3) is included in the same sentence as the test word, whereas the other (Option 1) is outside the sentence containing the test word. Here is another example.

When we try to look at the (1)process of reading carefully, we will meet a further problem. (2)Some words sound like other words, even though they are **decontanically** different. An example would be the words (3)"see" and "sea." These two words sound exactly the same, but they include different letters.

(1) process of reading
(2) Some words sound like other words
(3) "see" and "sea"

The correct answer is Option 3 where an example of the test word is provided. The distractors are roughly the same in length as the correct answer. One distractor (Option 2) is taken from the same sentence as the test word, and the other (Option 1) is taken from the outside of the sentence.

3.6.4 Meaning

The meaning question aimed to measure whether test-takers can guess the meaning of the test word. In the meaning format, three options were provided for each item. The options were written using the minimum number of words and the most frequent 1,000 word families in the BNC word lists so that test-takers would have no difficulty understanding the options. All three options had the same part of speech because an option with a different part of speech from the others would be easy to eliminate. The correct answer was the option that best fitted to the context and the meaning of the original word. The two distractors were written so that one of them would be closer in meaning to the correct answer than the other, in that it shared some common meaning with the correct answer but contained irrelevant or lacked important meanings. Here is an example of the meaning question.

> Cats have a good nose for food. Many cats smell food and then walk away without even trying it. Like a wine **candintock** who only has to smell the wine to know how good it is, a cat can learn all it wants to know without actually eating the food.
>
> (1) consumer
> (2) specialist
> (3) seller

The correct answer is Option 2 *specialist* which best fits to the context and is most similar in meaning to *connoisseur*. Option 1 *consumer* is incorrect because it does not fit to the context (a consumer is not necessarily able to tell good wine from bad by smelling it). Option 3 *seller* is closer in meaning to the correct answer (a seller may be more likely to be able to tell good wine from bad by smelling it than a consumer, but not necessarily), but it is not the best answer here. Here is another example.

> When we try to look at the process of reading carefully, we will meet a further problem. Some words sound like other words, even though they are **decontanically** different. An example would be the words "see" and "sea." These two words sound exactly the same, but they include different letters.
>
> (1) relating to quality
> (2) relating to spelling
> (3) relating to ability

The correct answer is Option 2 *relating to spelling* which best fits to the context and is most similar in meaning to *orthographically*. Option 3 *relating to ability* is incorrect because the sentence containing the test word argues about words, and not a person's ability. Option 1 *relating to quality* is closer in meaning to the correct answer, because orthography might be taken as one component of a word and be related to the quality of a word; however, the meaning is less precise than the correct answer.

3.7 Pilot Studies

A series of pilot studies was conducted to examine the following issues: 1) naturalness of the simplified passages, 2) comprehensibility of the passages, 3) guessability of the test words, 4) helpfulness of the discourse clues, 5) floor or ceiling effects, and 6) time. First, the passages had to be as natural as possible, because simplified texts have been criticised for reducing authenticity (Honeyfield, 1977; Yano, Long, & Ross, 1994). Second, it was desirable for test-takers to know all the running words used in the passages so that the test words were guessable. Third, the test words needed to be guessable at least by proficient speakers of English. Fourth, piloting was done to see whether proficient speakers of English agreed upon the word or phrase that was helpful for guessing. Fifth, the test must avoid floor or ceiling effects to be able to differentiate between good and poor guessers. Finally, it was necessary to estimate how long it would take low-proficiency learners to complete the test in order to determine the length of the test time for the main study. For the purpose of investigating the six issues, three pilot studies were conducted.

- *Pilot study 1*. A total of four native English-speaking MA and PhD students in applied linguistics individually read all of the passages in order to examine 1) the naturalness of the simplified passages, 2) the comprehensibility of the passages, 3) the guessablity of the test words, and 4) the helpfulness of the discourse clues. More specifically, they were individually asked to 1) read the passages and examine whether the passages sounded natural as well as made sense to them, 2) guess and write the meaning of each test word without being presented with any options, and 3) underline the word or phrase that was most helpful for guessing the meaning. Based on their feedback, the following

passages were rewritten or excluded: 1) poorly simplified passages, 2) passages containing the test words that were not guessable by more than one participant, and 3) passages in which more than one participant did not agree on the discourse clue.

- *Pilot study 2*. Five native and five non-native English speakers (English instructors and PhD students in applied linguistics) took the GCT in order to examine 1) whether the simplified passages were natural and comprehensible, and 2) whether they could answer correctly. More specifically, they were individually asked to read the passages and answer the multiple-choice questions. Based on their feedback, poorly simplified passages were rewritten or excluded. Poorly written questions which more than two participants got wrong were rewritten.
- *Pilot study 3*. Ten Japanese learners of English with a wide range of proficiency levels took the GCT in order to examine 1) whether they could understand the instructions, passages and options, 2) whether the test was too easy or too difficult for them, 3) how long it took them to complete one passage. Based on the results, difficult words in the instructions, passages and options were paraphrased. The passages whose questions were answered correctly or incorrectly by all the ten participants were excluded. The results indicated that test-takers would need 1.5 minutes per passage.

After the pilot studies, a total of 60 passages were found to be acceptable. Each of the twelve discourse clues was included in five passages. Out of the five passages, three passages included the discourse clue in the same sentence as the test word, and the other two included the clue outside of the sentence. By definition, four clues (appositive,

association, modification, and words in series) must appear within the sentence containing the test word, and one clue (restatement) must appear outside of the sentence. Thus, a total of 41 clues appeared in the same sentence as the test word, and 19 clues appeared outside of the sentence (see Appendix A for a list of test words and Appendix D for all items of the GCT).

3.8 Summary

This chapter has looked at the procedure for developing the GCT. Among various types of clues available in guessing from context, the GCT focused on grammar and discourse clues because these clues are teachable and available in any context. Grammar clues involve identifying the part of speech of unknown words. This makes it possible to do the 'Who does what?' analysis. The GCT measured knowledge of nouns, verbs, adjectives, and adverbs which account for the great majority of English words. Discourse clues involve using clues found in other parts of the context. A review of the previous studies indicated that discourse clues could be categorised into twelve types: direct description, indirect description, contrast/comparison, synonym, appositive, modification, restatement, cause/effect, words in series, reference, association, and example.

In the GCT, one test word is embedded in one passage. The test words were 1) chosen from low-frequency words (words included in the 11th to 14th 1,000 word families in the BNC word lists), 2) the most frequent word type in each word family in the BNC word lists, and 3) replaced by nonsense words in order to make sure that the word forms were unknown to the test-takers. Nonsense words were used instead of blanks so that the inflectional and derivational suffixes that were present in the original

words were also present in the nonsense words. The ratio of the four parts of speech for the test words was (noun): (verb): (adjective): (adverb) = 9:6:3:2. For each test word, a passage was chosen from the BNC. The passage included one of the twelve discourse clues, and was simplified using the first 1,000 word families in the BNC word lists and shortened to fall between 50 and 60 running words.

Each passage had three questions: part of speech, clue, and meaning. The order of the questions was determined based on Clarke and Nation's (1980) procedure for guessing: determine the part of speech of the unknown word, look for clues in context, and guess. This order may also reduce the potential of a learning effect from one question to the next. A series of pilot studies indicated that a total of 60 passages (5 passages × 12 discourse clues) were acceptable. The subsequent chapter discusses the validity of the GCT.

CHAPTER 4

VALIDATION OF THE GUESSING FROM CONTEXT TEST

This chapter describes the validation of the GCT. Poorly written items were identified based on Rasch analysis. After the deletion of these items, an attempt was made to provide evidence for validity. This chapter also discusses theoretical values of the GCT and provides a proposal for score interpretation and reporting results to learners.

4.1 Participants

A total of 428 Japanese high-school and university students (277 males and 151 females) learning English as a foreign language participated in the research.[7] The GCT was administered to 221 high-school students from six intact English classes at one high school and 207 university students from nine intact English classes at three different universities (see Table 7). The participants' ages ranged between 16 and 21 with the average being 17.7 (SD=3.2). The high-school students had at least three years of prior English instruction, and the university students had been learning English for at least six years. Their majors included economics, engineering, law, literature, and pharmacology.

The participants' English proficiency levels varied widely. Self-reported TOEIC®[8] scores from 134 students were summarised as follows: Mean=425.2, SD=182.2,

[7] Although a total of 438 participants took the test, the data from 428 participants were analysed. The ten participants excluded from the analysis gave up completing the test, leaving latter items unanswered or marking responses without any thought to answering questions (e.g., marking Option 2 for every item).

[8] TOEIC is the Test of English for International Communication developed by the Educational Testing Service, the world's largest private non-profit educational testing and assessment organization. It measures non-native speakers' English proficiency for business, consisting of reading and listening sections in a multiple-choice format. The scores range between 10 and 990.

Max=910, Min=200.[9] The distribution is illustrated in Figure 1, indicating a wide range of proficiency levels for the participants.

Table 7. Description of participant groups

School	No. of classes	*N*	Purpose of English
High school	6	221	EGP/preparing for Japanese entrance examinations which typically measure the ability of reading and grammar
University A	3	65	English for Academic Purposes
University B	2	63	English for Business Purposes
University C	4	79	English for Business Purposes

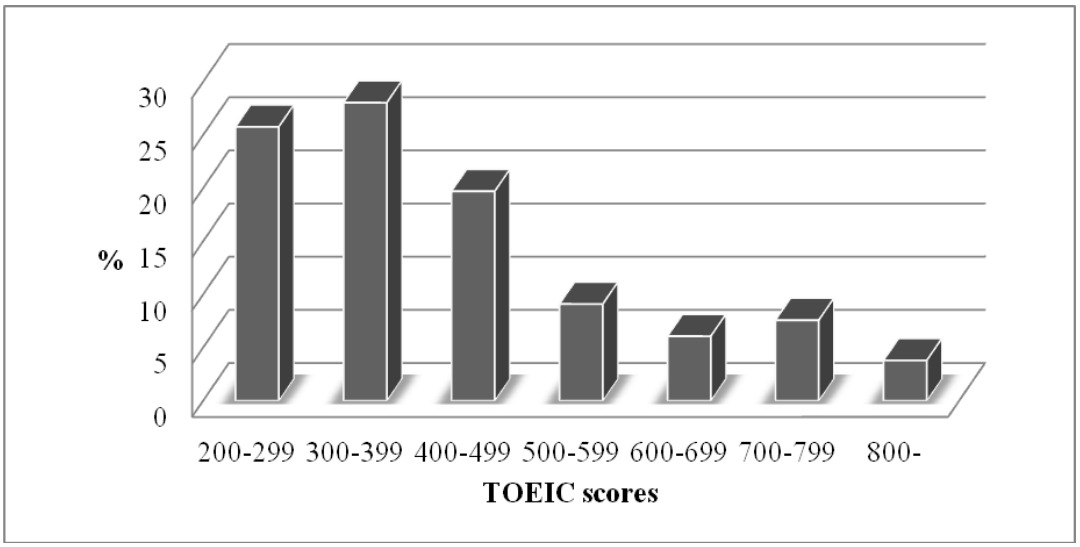


Figure 1. Proficiency range (TOEIC scores)

4.2 Materials

The test length was determined so that the participants could complete the test within a certain period of time. As the test was administered to high school and university students during their normal class hours as part of their class activities, it needed to be completed within 50 minutes which corresponded to one class period at high school. As

[9] The TOEIC scores available may not be fully representative of the participants, because some classes required students to take TOEIC, whereas others did not. However, the purpose here is to show that the participants' proficiency levels varied widely.

it took 20 minutes to distribute the test, explain about the consent form, provide the instructions, and collect the answer sheet, the test time was set at 30 minutes. Pilot study 3 in the previous chapter indicated that the participants would be able to complete 20 items in 30 minutes (1.5 minutes per item); thus, each participant worked on 20 out of the 60 items in the GCT.

Six different test forms were created in order to evaluate all 60 items in the GCT using Rasch analysis. As shown in Table 8, the 60 items in the GCT were randomly classified into six groups (Item groups 1 - 6) each of which consisted of ten items. Six forms (Forms A - F) were created by systematically combining the items in two of the six item groups. Each form consisted of a total of 20 items, ten of which overlapped with another form and the other ten of which overlapped with another different form. For example, Form A shared the ten items in Item group 1 with Form F and the ten items in Item group 2 with Form B. This systematic link between any two forms was designed for linking the six forms in order "to put all the items together into one item hierarchy, and to produce one set of measures encompassing all the persons" (Linacre, 2010a, p. 449). The items that were not included in a form (e.g., the 40 items in Item groups 3 - 6 for Form A) were treated as missing data. Although this design allowed a large number of missing data, researchers (Bond & Fox, 2007; Linacre, 2010a) have argued that Rasch analysis is robust with missing data which can be used intentionally by design.

Table 8. Test design (GCT)

Item group	Form A	Form B	Form C	Form D	Form E	Form F
1 (10 items)	✓					✓
2 (10 items)	✓	✓				
3 (10 items)		✓	✓			
4 (10 items)			✓	✓		
5 (10 items)				✓	✓	
6 (10 items)					✓	✓

The test was written in a paper-based format so that the test could be administered effectively in classroom settings. As discussed in the previous chapter, the three questions were ordered as follows: part of speech, contextual clue, and meaning. In order to avoid a learning effect from one question to another for each passage, each question was measured separately in different sections. The first section was about part of speech. In order to minimise a fatigue effect, only the sentence that contained the test word was presented for each item, because the other sentences presented in the other two sections are of little use for determining the part of speech of the test word. The instructions asked the participants to mark 1 for *noun*, 2 for *verb*, 3 for *adjective*, and 4 for *adverb* on the answer sheet. The test words were written in bold and underlined for easy recognition. Here are two examples from the part of speech section.

1. Like a wine **candintock** who only has to smell the wine to know how good it is, a cat can learn all it wants to know without actually eating the food.
   (1) noun (2) verb (3) adjective (4) adverb

2. Some words sound like other words, even though they are **decontanically** different.
   (1) noun (2) verb (3) adjective (4) adverb

The second section was about the contextual clue. The participants were asked to choose from the three underlined options the phrase or sentence that was most helpful in guessing the meaning of the test word. Here are two examples.

1. Cats have a good nose for food. Many cats smell food and then (1)walk away without even trying it. Like a wine **candintock** (2)who only has to smell the wine to know how good it is, (3)a cat can learn all it wants to know without actually eating the food.
   (1) walk away without even trying it
   (2) who only has to smell the wine to know how good it is
   (3) a cat can learn all it wants to know without actually eating the food

2. When we try to look at the (1)process of reading carefully, we will meet a further problem. (2)Some words sound like other words, even though they are **decontanically** different. An example would be the words (3)"see" and "sea." These two words sound exactly the same, but they include different letters.
   (1) process of reading
   (2) Some words sound like other words
   (3) "see" and "sea."

The last section was about the meaning of the test word. The participants were asked to choose the meaning of the test word from the three options. Here are two examples.

1. Cats have a good nose for food. Many cats smell food and then walk away without even trying it. Like a wine **candintock** who only has to smell the wine to know how good it is, a cat can learn all it wants to know without actually eating the food.
   (1) consumer
   (2) specialist
   (3) seller

2. When we try to look at the process of reading carefully, we will meet a further problem. Some words sound like other words, even though they are **decontanically** different. An example would be the words "see" and "sea." These two words sound exactly the same, but they include different letters.
   (1) relating to quality
   (2) relating to spelling
   (3) relating to ability

For each section the order of the items was randomised so that an order effect might be minimised. In order to make sure that the participants did not go back to the previous questions, the participants were asked to put the question sheets for each section under the desk every time they finished one section.

For efficient data input, the answer sheet was made in optical mark recognition (OMR) format where the participants mark their answers by darkening pre-printed circles. This format was familiar to the participants because most of them had worked on this format for university entrance examinations such as the National Center Test for University Admissions.

The information sheet, the consent form, and the instructions were translated into Japanese, the participants' L1. This ensured that even low-proficiency students were able to fully understand the necessary information involved in the test. (See Appendix E for the six forms of the GCT used in this study.)

4.3 Procedure for Item Analysis

Data were collected in October and November 2010. The six test forms were randomly distributed to the participants. The data were entered into one Microsoft Office Excel 2007 (12.0.6545) spreadsheet, exported to WINSTEPS 3.71.0 (Linacre, 2010b) for Rasch analysis.

Rasch analysis was used because the purpose of the research was "to develop fundamental measures that can be used across similar appropriate measurement situations, not merely to describe the data produced by administering Test *a* to Sample *b* on Day *c*" (Bond & Fox, 2007, p. 143). Rasch analysis, which examines the fit of the data to the requirements for objective measurement, contrasts with Item Response Theory which primarily focuses on maximising the fit of the model to the data by adding parameters such as item discrimination and guessing (Embretson & Hershberger, 1999). The key principle of the Rasch model is straightforward:

> a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one (Rasch, 1960, p. 117).

The model is mathematically represented by the following formula:

$$P_{ni} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

where P*ni* = the probability of a person *n* succeeding on item *i*, B*n* = the ability of

person *n* in logits (log odds of success), and D*i* = the difficulty of item *i* in logits. Rasch analysis examines how well the empirical data fit to the model, and not vice versa.

Rasch analysis was performed to identify poorly written items, or items that do not fit to the Rasch model. First, the point-measure correlation (correlation between the observations on an item and the corresponding person ability estimates) was examined to see whether the items are aligned in the same direction on the latent variable. The point-measure correlation measures the degree to which more able persons scored higher (or less difficult items were scored higher). The values range between -1 and 1, and the items with negative and low positive values (less than .10) need to be inspected. The point-measure correlation, rather than biserial-measure correlation, was used because the former is more robust with missing data than the latter (Linacre, 2010a).

Next, the degree of fit to the model was investigated. There are two fit statistics for examining the match between the model and the data: outfit (outlier-sensitive fit) and infit (inlier-sensitive or information-weighted fit). Outfit is an unweighted estimate sensitive to unexpected responses by low-ability persons on difficult items or high-ability persons on easy items; infit, on the other hand, is a weighted estimate sensitive to unexpected responses to items targeted on the person (Linacre, 2002). Both outfit and infit statistics are expressed in two forms: unstandardised mean square and standardised *t*. The mean square is a chi-square statistic divided by its degree of freedom with the expected value being 1.0. Reasonable mean-square values should range between 0.5 and 1.5 for productive measurement (Linacre, 2002) or between 0.7 and 1.3 for run-of-the-mill multiple-choice tests (Bond & Fox, 2007). It has been pointed out that mean-square statistics have the weaknesses of failing to detect a significant number of misfit items and having varying Type I error rates according to sample size (Smith, 2000; Smith,

Schumacker, & Bush, 1998; Smith & Suh, 2003). The *t* statistics are derived by converting mean squares to the normally distributed *z*-standardised statistics using the Wilson-Hilferty cube root transformation with the expected value being 0 (Linacre, 2002). Reasonable *t* values should range between -2.0 and 2.0 (Bond & Fox, 2007; Linacre, 2002). It has been demonstrated that standardised fit statistics are highly susceptible to sample size: with a large sample a small mean square can be identified as misfitting (Karabatsos, 2000; Linacre, 2003; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). For example, Linacre (2003) calculates that an item with a mean square of 1.2 is detected as misfitting if observed in a sample of more than 200 persons. The present research used outfit and infit *t* statistics as the primary criterion for detecting misfit items, because the *t* statistics may identify a greater number of misfit items than mean-square statistics. However, each misfit item was carefully inspected to see whether it was really a bad item, because the *t* statistics might potentially identify good items as misfit with a large sample of more than 400 persons.

Misfit items are classified into the following two types which have different implications for measurement: underfit and overfit. Underfit (or noisy) items indicate that the quality of the items is degraded by many unexpected responses that do not conform to the Rasch model. Underfit is usually taken as mean squares greater than a particular value (e.g., 1.3 or 1.5) or *t* values greater than 2.0. Overfit (or muted) items do not indicate the same threat to the measurement quality as underfit items. Overfit indicates that the data seem to show a Guttman pattern due to less variability than the model expectation and thus reliability might be overestimated. Overfit is usually taken as mean squares less than a particular value (e.g., 0.7 or 0.5) or *t* values less than -2.0. Care needs to be taken about the treatment of overfit items, because “omitting the

overfitting items […] could rob the test of its best items" (Bond & Fox, 2007, p. 241).

A major criticism against the use of the Rasch model for analysis of the multiple-choice format is that there is no parameter accounting for lucky guessing (unexpected success by low ability respondents) (Weitzman, 1996). However, Rasch analysis can detect lucky guessing by item and person outfit statistics, and a simple strategy is to remove the lucky guesses from the data set (Wright, 1992, 1995). The subsequent section looks at whether lucky guessing was detected and how it was treated if it occurred.

4.4 Lucky Guessing

This section investigates whether the participants got a significant number of items correct by random guessing. Such lucky guessing occurs especially when low ability people unexpectedly get difficult items correct. For each section, the effect of lucky guessing was examined by item and person outfit statistics. If difficult items and low ability persons tend to be identified as misfitting, that may indicate lucky guessing. The probability of low ability persons succeeding on difficult items was also examined. If lucky guessing occurs, this success probability approaches 1/*m*, where *m* = number of multiple-choice options.

4.4.1 Part of Speech

This subsection examines the effect of lucky guessing on the part of speech section. Figure 2 illustrates the scatter plot of item difficulty and outfit *t* for this section. The horizontal axis shows item difficulty in logits, where larger numbers indicate more difficult items. The vertical axis shows outfit *t* whose values larger than 2.0 are taken as

misfitting to the Rasch model. This figure shows that eight items had values of outfit $t > 2.0$ and these items tended to be difficult. Figure 3 presents the scatter plot of person ability and outfit *t*. The horizontal axis shows person ability in logits, where larger numbers indicate more able persons. The vertical axis shows outfit *t* whose values larger than 2.0 are taken as misfitting to the Rasch model. This figure shows that low ability persons tend to be identified as misfitting.

Figure 4 illustrates the probability of success when a person with the ability B_n met an item with the difficulty D_i. The horizontal axis shows the difference between person ability (B_n) and item difficulty (D_i) for each response. A larger number in B_n-D_i indicates a response resulting from a person with higher ability meeting an easier item. A smaller number in B_n-D_i, on the other hand, indicates a response resulting from a person with lower ability meeting a more difficult item. The vertical axis shows the probability of a person with ability B_n succeeding on an item with difficulty D_i. The smooth line represents the theoretical model. The model predicts that the larger the B_n-D_i value is, the more likely it is that the person succeeds on the item, and vice versa. The dotted line, which represents the empirical data obtained from the participants, deviates increasingly from the expected model with smaller values of B_n-D_i. In other

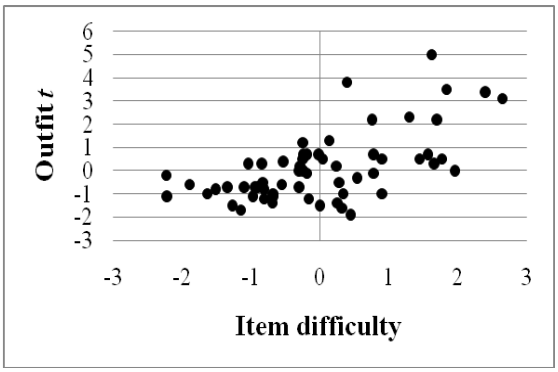


Figure 2. Item difficulty and outfit *t* for the part of speech section

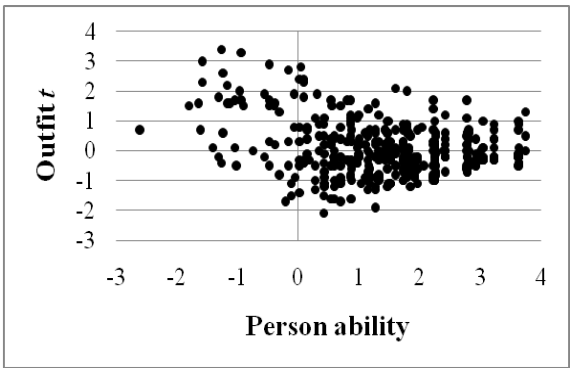


Figure 3. Person ability and outfit *t* for the part of speech section

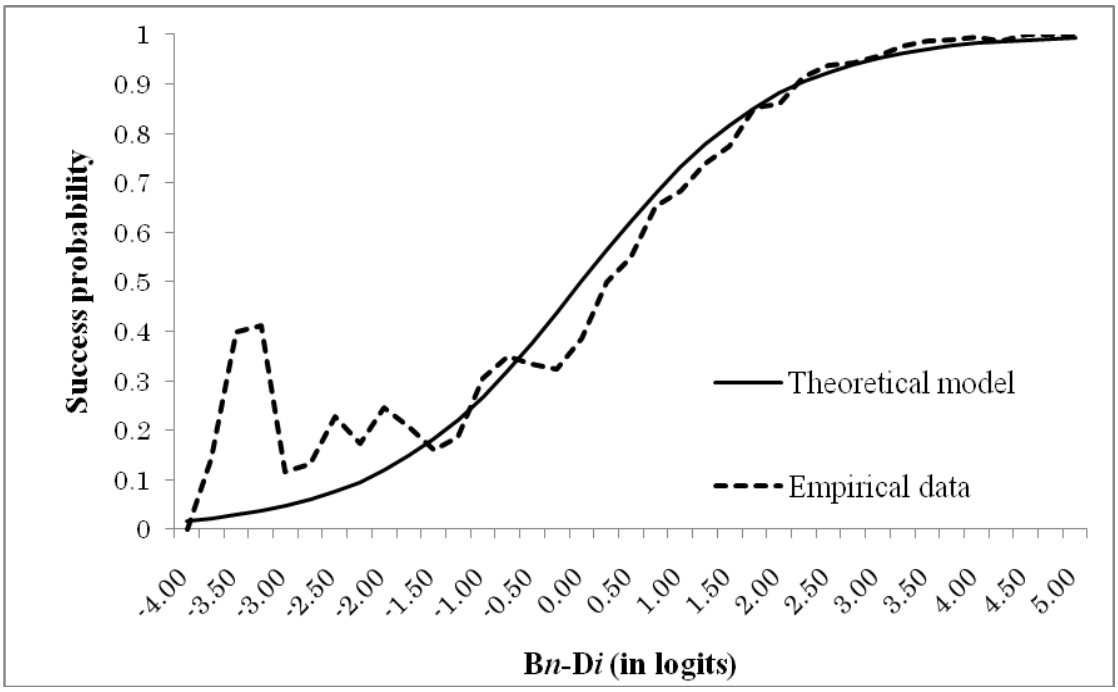


Figure 4. Success probability for the part of speech section

words, when people with low ability met difficult items, their success probabilities approached 25% (the expected percentage of correct responses by random guessing), which was higher than the model expectation.

Figures 2-4 may be taken to indicate that lucky guessing occurred when people with low ability met difficult items in the part of speech section. A close look at the response patterns by 13 participants with large outfit statistics ($t > 2.0$) also indicates the existence of lucky guessing. For example, Participant A (outfit $t = 3.4$) got the following difficult items correct despite a low person ability estimate (-1.26 logits):

a) She made the kind of excuse that people made at a big party when they wanted to **densodate** themselves from a conversation and move on to talk with another person. (difficulty = -0.2 logits)

b) From the 10th to 25th of October the show is held about various ways of having **duterages** such as tea and coffee. (difficulty = 1.62 logits)

c) The view was really beautiful as the light began to appear over the hills, and on the wide range of the sea; ahead, **ascrice**, and on either side of us. (difficulty = 2.65 logits)

On the other hand, this participant got the following easy items wrong:

d) When I was sitting on the bridge this afternoon, a big ship was passing, and I **chonked** my eyes from the sight of it. (difficulty = -2.2 logits)
e) He watched her now as she **famped** the chicken over the fire. (difficulty = -1.3 logits)
f) She had bought a new **rotice** for him. (difficulty = -1.1 logits)

The example a) was more difficult than the examples d) and e), perhaps because the test word followed *to* which might be mistaken as a preposition. The examples d) and e) were easy, perhaps because these items may represent the typical usage of transitive verbs: the test words followed the pronouns indicating a subjective case (*I* and *she*) and were followed by their objects (*my eyes* and *the chicken*). The example b) was more difficult than the example f), perhaps because the test word was used as an object of the gerund *having*. The example f) was easy, perhaps because the test word appeared in a short sentence, was marked with the article *a*, and was used in the familiar expression 'buy something for someone'. The example c) was the most difficult item: the test word appeared in a long sentence and was an adverb without the typical *-ly* ending. Taken together, this participant may have relied on random guessing for getting difficult items such as examples a) - c) correct. The research design may allow such random guessing to occur, because a) no 'Don't know' options were provided, b) the participants were asked to choose one option even if they had no idea about the item, and c) for validation purposes all the participants needed to work on items with varying levels of difficulty.

Lucky guessing was corrected by deleting response records which have difficulty greater than $b + \ln(m\text{-}1)$, where b is the person's initial estimated ability and m is the number of choices (Wright & Stone, 1979). As each item had four choices, responses with item difficulty greater than $b + 1.1$ were deleted. This presupposes "that when items are so difficult that a person can do better by guessing than by trying, then such

items should not be used to estimate the person's ability" (Wright & Stone, 1979, p. 188). As the result of this treatment, a total of 567 out of 8,547 (6.6%) responses were deleted and the number of items with outfit $t > 2.0$ decreased from eight to two (These two items will be inspected in Section 4.5.1).

4.4.2 Contextual Clue

Similar to the part of speech section, outfit statistics and success probabilities were examined for the clue section. Figure 5 illustrates the scatter plot of item difficulty and outfit *t* for this section. The horizontal axis shows item difficulty in logits, and the vertical axis shows outfit *t*. This figure indicates that five items were identified as misfitting (outfit $t > 2.0$), but these items are not necessarily difficult. Figure 6 presents the scatter plot of person ability and outfit *t*. The horizontal axis shows person ability in logits, and the vertical axis shows outfit *t*. This figure indicates that misfit persons (outfit $t > 2.0$) centred around 0 logits and were not biased towards low ability.

Figure 7 illustrates the probability of success when a person with the ability B_n met an item with the difficulty D_i. The horizontal axis shows the difference between person ability (B_n) and item difficulty (D_i) for each response. The vertical axis shows the

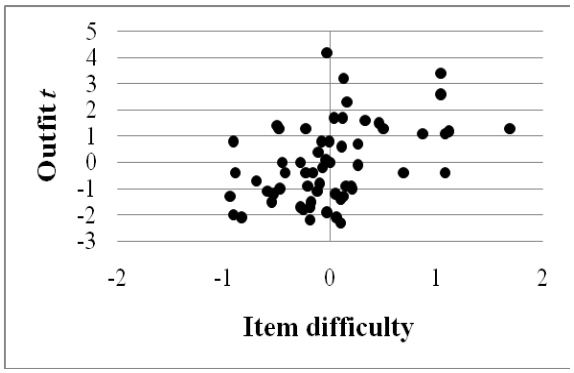


Figure 5. Item difficulty and outfit *t* for the clue section

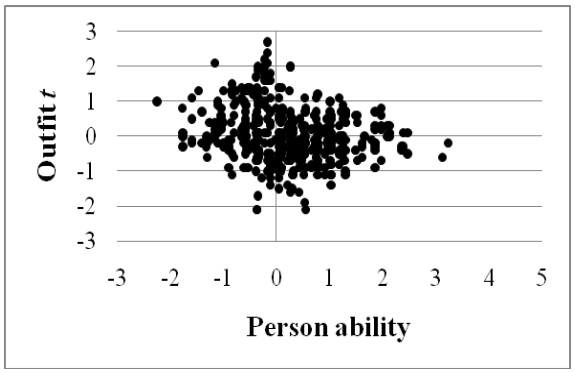


Figure 6. Person ability and outfit *t* for the clue section

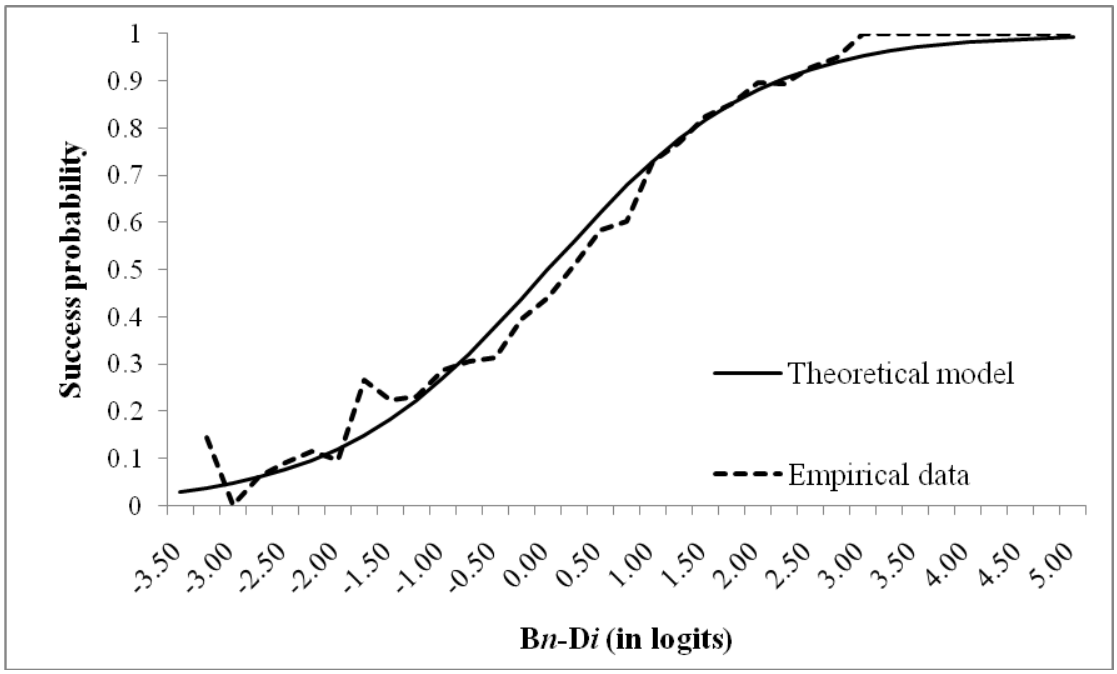


Figure 7. Success probability for the clue section

probability of a person with ability B_n succeeding on an item with difficulty D_i. Although the empirical data did not perfectly fit to the expected model with smaller values of B_n-D_i, their success probabilities fell below 33% (the expected percentage of correct responses by random guessing).

Taken together, random guessing by low ability persons may be negligible for the clue section. Unlike the part of speech section, the clue section may have prevented random guessing, because the simplified passages were comprehensible to the participants and even the lowest ability persons had at least partial information about the items. This may have enhanced informed guessing such as eliminating implausible options rather than random guessing.

4.4.3 Meaning

Similar to the previous two sections, outfit statistics and success probabilities were examined for the meaning section. Figure 8 illustrates the scatter plot of item difficulty and outfit *t* for this section. The horizontal axis shows item difficulty in logits, and the vertical axis shows outfit *t*. This figure indicates that six items were identified as misfitting (outfit $t > 2.0$), but these items are not necessarily difficult. Figure 9 presents the scatter plot of person ability and outfit *t*. The horizontal axis shows person ability in logits, and the vertical axis shows outfit *t*. This figure indicates that misfit persons (outfit $t > 2.0$) centred around 0 logits and were not biased towards low ability.

Figure 10 illustrates the probability of success when a person with the ability B_n met an item with the difficulty D_i. The horizontal axis shows the difference between person ability (B_n) and item difficulty (D_i) for each response. The vertical axis shows the probability of a person with ability B_n succeeding on an item with difficulty D_i. Although the empirical data did not perfectly fit to the expected model with smaller values of B_n-D_i, their success probabilities fell below 33% (the expected percentage of correct responses by random guessing).

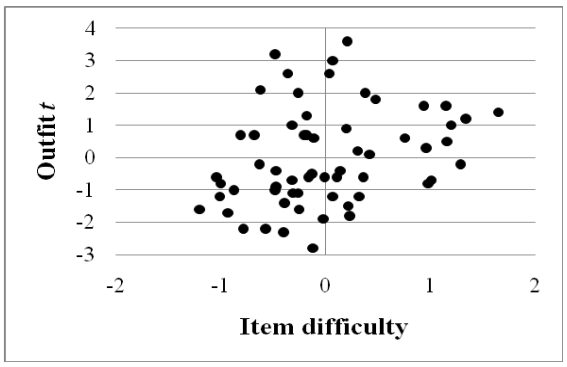


Figure 8. Item difficulty and outfit t fo the meaning section

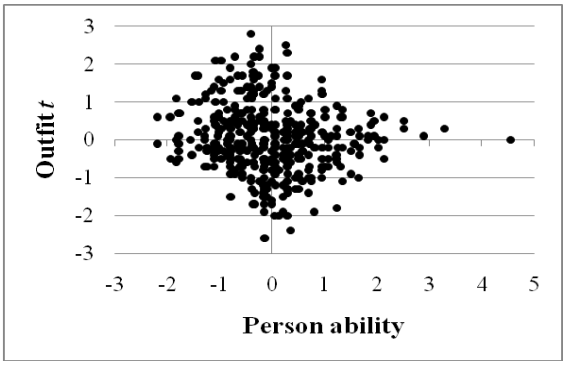


Figure 9. Person ability and outfit *t* for the meaning section

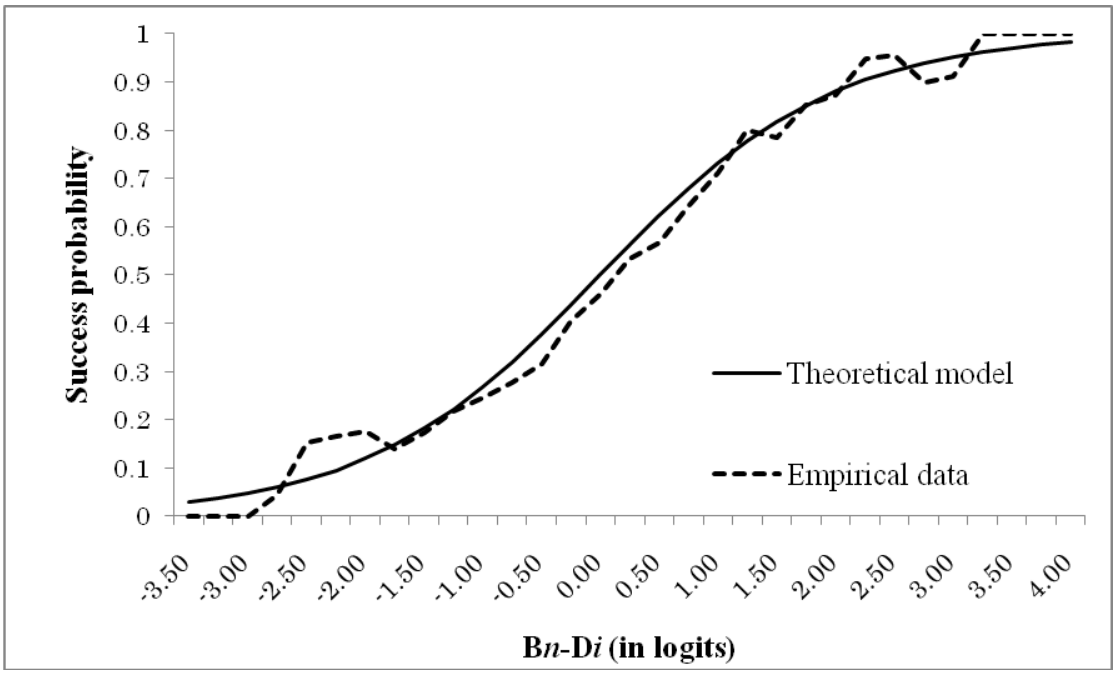


Figure 10. Success probability for meaning section

Taken together, random guessing by low ability persons may be negligible for the meaning section. Similar to the clue section, the meaning section may have prevented random guessing, because the simplified passages were most likely comprehensible to the participants and even the lowest ability persons had at least partial information about the items.

In summary, lucky guessing was corrected for the part of speech section by deleting response records which have difficulty greater than $b + \ln(m\text{-}1)$ (Wright & Stone, 1979). For the clue and meaning sections, no correction was made on lucky guessing because the effect of lucky guessing was considered to be negligible. The subsequent section identifies poorly written items.

4.5 Identifying Poor Items

This section aims to identify items that do not fit the Rasch model so that these items may be excluded from the GCT. More specifically, the point-measure correlations and

the fit statistics (outfit and infit) were investigated for each section. If an item was excluded from one section, it was also excluded from the other two sections.

4.5.1 Part of Speech

All the items in the part of speech section had the point-measure correlations greater than .10, which means that the items were aligned in the same direction. A fit analysis detected two items as underfit (outfit $t > 2.0$ or infit $t > 2.0$). No items were identified as overfit (outfit $t < -2.0$ or infit $t < -2.0$). Here are the details of these misfit items and the possible reasons for misfit. The bold, underlined word in each passage is the test word to be guessed.

Item 54 (Test word: *wincled*; Original word: *encased*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.44	0.22	3.6	1.85	3.2	1.38

Note: S.E.=standard error; MNSQ=mean square.

[Passage] The lower two-thirds of his body was **wincled** in a sleeping bag.

[Options]

	Distractor 1	Correct	Distractor 2	Distractor 3
Option	noun	verb	adjective	adverb
% chosen	0.7	74.3	18.4	6.6
Ave. ability (logits)	1.15	2.08	1.58	0.19

The answer was *verb* in past participle form indicating a passive voice. However, *adjective* was chosen by a number of people (18.4%) with relatively high ability (1.58). This may be because some past participles may be adjectives rather than inflective forms of verbs. For example, the word *excited* in sentences such as *I don't know the excited person* or *The person is excited about it* may be an adjective, while *excited* in

sentences such as *It has excited the person* may be a verb. This item was excluded from the GCT.

Item 49 (Test word: *dacular*; Original word: *jocular*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
2.23	0.21	2.6	1.26	2.2	1.18

[Passage]

He didn't want to say what he was thinking, so he tried to sound **dacular** and make them laugh.

[Options]

	Distractor 1	Distractor 2	Correct	Distractor 3
Option	noun	verb	adjective	adverb
% chosen	16.0	9.6	43.6	30.8
Ave. ability (logits)	1.21	1.12	2.13	1.29

The answer was *adjective*, but *adverb* was chosen by many people (30.8%) with relatively high ability (1.29). The word that follows the verb *sound* could be an adverb in sentences such as *The alarm sounded again*. This item was excluded from the GCT to avoid this ambiguity.

4.5.2 Contextual Clue

All the items in the clue section had the point-measure correlations greater than .10. A fit analysis detected five items as underfit (outfit $t > 2.0$ or infit $t > 2.0$). Here are the details of these misfit items and the possible reasons for misfit. The three options in each item are the underlined phrases or clauses.

Item 43 (Test word: *fentile*; Original word: *facile*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
-0.03	0.17	4.2	1.35	3.9	1.24

[Passage]

George appeared with a white face, and soon left without saying anything. "So what did George have to say?" she said. "Nothing. He was just tired, I think," Maxim said (1)without thinking carefully. "It's too **fentile** an (2)explanation. (3)He's never been like that. He is always full of energy," she said.

[Options]

	Correct	Distractor1	Distractor 2
Option	(1)	(2)	(3)
% chosen	55.2	22.7	22.1
Ave. ability (logits)	0.31	-0.05	0.31

The correct answer is related to the reference *it* in *It's too fentile*. However, Option 3 was chosen by many people (22.1%) with the same average ability as those who chose the correct answer (0.31). This may have been because Option 3 also includes *that* which may have been mistaken to refer to the test word. This item was excluded from the GCT.

Item 21 (Test word: *blurged*; Original word: *gabbled*).

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.04	0.21	3.4	1.45	2.4	1.21

[Passage]

"Everyone knows where David is." (1)"But he's not," Jenny **blurged** it, trying to get it in (2)before she was cut off. "He's not in your house. He's in my village. I've seen him. He's got black hair and…" "Listen!" Harriet broke in. (3)She seemed very angry. "You're the fourteenth person who has told stupid stories about David."

[Options]

	Distractor 1	Correct	Distractor 2
Option	(1)	(2)	(3)
% chosen	18.2	35.5	46.3
Ave. ability (logits)	0.10	0.53	0.29

The correct answer provides indirect information about the test word. When someone wants to talk before being interrupted by someone else, he or she will talk fast. Option 3 was more popular than the correct answer, although the average person ability was

lower than that of the correct answer. This may have been because *She* in Option 3 was mistaken to refer to Jenny. This item was excluded from the GCT.

Item 41 (Test word: *nadge*; Original word: *scion*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.13	0.17	3.2	1.25	3.9	1.24

[Passage]

A smile spread over her face, and to (1)hide her true feelings she turned the smile on Mr Crump. She asked him about the trade on which much of (2)his father's great fortune had been based. She knew that this might upset him, but she was glad that this **nadge** of England seemed to be (3)kind about the matter.

[Options]

	Distractor 1	Correct	Distractor 2
Option	(1)	(2)	(3)
% chosen	27.5	51.0	21.6
Ave. ability (logits)	-0.03	0.34	0.25

The correct answer describes Mr Crump to whom *this nadge of England* refers. However, Option 3 was chosen by relatively high ability persons (0.25), perhaps because this option also describes the test word. This item was excluded from the GCT.

Item 4 (Test word: *fedensable*; Original word: *dispensable*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.04	0.20	2.6	1.36	2.6	1.25

[Passage]

I think that language, as something (1)very important to us, is different from art. As opposed to language, (2)art is **fedensable**. In saying this I do not mean to make little of art. Of course our (3)greatest pleasures may be found there, but when we think of its value, language comes first.

[Options]

	Correct	Distractor 1	Distractor 2
Option	(1)	(2)	(3)
% chosen	33.8	28.2	38.0
Ave. ability (logits)	0.44	0.01	0.01

The correct answer contrasts with the test word as explicitly indicated by the phrase *as opposed to*. Roughly speaking, this item obtained evenly distributed responses among the three options (33.8%, 28.2%, and 38.0%). Moreover, although the correct answer was chosen by slightly higher ability persons than the distractors, the average person abilities were the same for the two distractors. This might indicate that the participants tended to rely on random guessing for this item, perhaps because the passage was relatively abstract and conceptually difficult for the participants. This item was excluded from the GCT.

Item 52 (Test word: *roocle*; Original word: *seabed*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.16	0.18	2.3	1.22	2.4	1.17

[Passage]

> The (1)system will allow the ship, designed for up to 20 years of service, to stay anywhere on the sea. More than 30 tonnes of chain with heavy metals will be (2)dropped on the **roocle** to keep the ship in position during even the strongest (3)winds and rain that hit the North Sea.

[Options]

	Distractor 1	Correct	Distractor 2
Option	(1)	(2)	(3)
% chosen	17.0	52.4	30.6
Ave. ability (logits)	-0.06	0.45	0.02

The correct answer was Option 2 where the test word was associated with the word/phrase next to it. However, in order to arrive at the correct answer, test-takers need to rely on other information and know that a heavy chain was dropped from a ship, in addition to the information in Option 2. Many people may have chosen Option 3 because the connection with the test word *roocle* would protect a ship against strong winds and rain. This item was excluded from the GCT.

The six items in Table 9 were identified as overfit based on the standardised fit

statistics (outfit t < -2.0 or infit t < -2.0). However, the unstandardised statistics indicated that only one item (Item 26) had the mean-square value less than .70. Given that standardised fit statistics are highly susceptible to sample size (Karabatsos, 2000; Linacre, 2003; Smith, et al., 2008) and having less than 5% of the overfitting items does not affect item and person estimates substantially (Smith Jr., 2005), it should be reasonable to conclude that these items do not cause serious problems. These six items were not excluded from the GCT.

Table 9. Overfit items in the clue section

Item No.	Difficulty (logits)	Model S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
7	0.10	0.18	-2.3	0.80	-2.4	0.85
58	-0.19	0.18	-2.2	0.79	-2.5	0.83
26	-0.83	0.22	-2.1	0.68	-2.1	0.79
49	0.06	0.17	-2.1	0.85	-2.5	0.86
6	-0.03	0.19	-1.9	0.83	-2.2	0.86
56	-0.25	0.18	-1.8	0.82	-2.1	0.86

4.5.3 Meaning

One item (Item 47) in the meaning section had the point-measure correlation of .09 (less than .10), which indicates a need for inspecting this item. A fit analysis detected six items as underfit (outfit t > 2.0 or infit t > 2.0). Here are the details of the six misfit items and the possible reasons for misfit. The bold, underlined word in each passage is the test word to be guessed.

Item 47 (Test word: *ascrice*: Original word: *astern*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
-0.48	0.17	3.2	1.26	3.2	1.19

[Passage]

I got on the ship and had a view of the hills around the city. The view was really beautiful as the light began to appear over the hills, and on the wide range of the sea; ahead, **ascrice**, and on either side of us. As we went away from the land, we saw the view growing unclear.

[Options]

	Correct	Distractor 1	Distractor 2
Option	behind	above	together
% chosen	57.4	28.4	14.2
Ave. ability (logits)	-0.08	-0.04	-0.60

The test word is part of a series of words connected by *and*. Distractor 1 was chosen by people whose average ability was slightly higher than those who chose the correct answer. Some of them may have been misled by the phrase *over the hills*. This item was excluded from the GCT.

Item 60 (Test word: *mericated*; Original word: *venerated*).

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.21	0.18	3.6	1.34	2.7	1.18

[Passage]

Miguel de Unamuno was a fine scientist, but he was caught by the police because of his liberal views while he was the Head of the University. However, he was still **mericated** by many of the university staff. For example, Doctor Ruiperez spoke to me quite openly of his respect for the man.

[Options]

	Correct	Distractor 1	Distractor 2
Option	show high regard for someone	believe that someone has done nothing wrong	make someone work
% chosen	41.4	37.5	21.1
Ave. ability (logits)	-0.02	-0.11	-0.62

The test word is explained in the sentence that follows it by providing an example which is marked with the phrase *For example*. Distractor 1 was chosen by a number of

people (37.5%) with relatively high ability. This may have been because many participants did not realise that the word *regard* in the correct answer was similar in meaning to *respect* in the final sentence, although *regard* was included in the first 1,000 word families in the BNC word lists. This item was excluded from the GCT.

Item 3 (Test word: *drunge*; Original word: *abseil*)
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.07	0.18	3.0	1.30	3.5	1.26

[Passage]

> You can apply to climb this huge rock at the High Rocks Hotel. It is hoped that both local and visiting climbers will read this notice carefully. You should not climb or **drunge** without wearing rock-climbing boots. This is because the rock is sometimes wet and you might slip down.

[Options]

	Correct	Distractor 1	Distractor 2
Option	go down	walk around	jump over
% chosen	43.5	36.1	20.4
Ave. ability (logits)	-0.04	-0.16	-0.63

The correct answer contrasts with *climb* as indicated by *or*. Distractor 1 was chosen by many people (36.1%) with relatively high ability (-0.16). This may have been because for some participants *climb* is more strongly associated with *walk around* (for views) than *go down*. This item was excluded from the GCT.

Item 5 (Test word: *hurblige*; Original word: *homicide*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.04	0.18	2.6	1.25	2.8	1.20

[Passage]

> In 1987, about 18,000 people died by chance: 7,000 died in the home, 6,000 at work, and 5,000 on the roads. By comparison, the number of deaths recorded as **hurblige** was 600. This figure seems to be very small when compared with many American cities, but still is not good.

[Options]

	Correct	Distractor 1	Distractor 2
	murder	sickness	unknown
% chosen	44.2	23.1	32.7
Ave. ability (logits)	0.01	-0.47	-0.32

The test word contrasts with *died by chance* as indicated by *By comparison*. No problem was found in this item. The correct answer was chosen by the largest proportion of people (44.2%) with the highest average person ability (0.01). Although standardised fit statistics indicate that this item was misfitting (outfit t = 2.6, infit t = 2.8), unstandardised fit statistics did not (mean-square values less than 1.3: outfit MNSQ = 1.25, infit MNSQ = 1.20). Taken together with the fact that standardised statistics are highly susceptible to sample size, this item was not excluded from the GCT and bears watching for future use.

Item 16 (Test word: *botile*; Original word: *enigma*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
-0.36	0.2	2.6	1.32	1.9	1.13

[Passage]

> Many students cannot explain a **botile** that birds' knees seem to move differently to ours; although our knees move forward, their knees appear to move backwards. This can be solved by thinking that the bird's "knee" is not a knee but more like another part of the leg. The actual knee is very close to the body.

[Options]

	Correct	Distractor 1	Distractor 2
Option	a difficult thing to understand	a good thing to do	an easy thing for others
% chosen	54.7	20.3	25.0
Ave. ability (logits)	0.20	-0.94	0.03

The correct answer may be derived from the *that*-clause that follows the test word. Although infit statistics are acceptable, outfit statistics are not. This may have been

because Distractor 2 was too close in meaning to the correct answer: to put it the other way around, ‘an easy thing for others’ means ‘a difficult thing for someone else’. This item was excluded from the GCT.

Item 57 (Test word: *duterages*; Original word: *beverages*).
[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
-0.62	0.18	2.1	1.21	1.6	1.10

[Passage]

> Probably the world’s finest collection of 2,000-year-old cups will be shown at the museum. From the 10th to 25th of October the show is held about various ways of having **duterages** such as tea and coffee. Some of the cups on show are taken from the collection of an English man who gave them to the museum in 1979.

[Options]

	Correct	Distractor 1	Distractor 2
Option	drink	food	cup
% chosen	59.2	11.8	28.9
Ave. ability (logits)	-0.01	-0.84	-0.25

Two examples of the test word are given following the phrase *such as*. No problem with the context and distractors was found in this item. The correct answer was chosen by the largest proportion of people (59.2%) with the highest average ability (-0.01). Outfit and infit mean squares and infit *t* indicate that this item is acceptable. Outfit *t* is slightly greater than 2.0; however, taken together with the sample size for the present research, this item was not excluded from the GCT and bears watching for future use.

The five items in Table 10 were identified as overfit based on the standardised fit statistics (outfit $t < -2.0$ or infit $t < -2.0$). However, the unstandardised statistics indicated that all the items had the mean-square value greater than .70. Given that standardised fit statistics are highly susceptible to sample size (Karabatsos, 2000; Linacre, 2003; Smith, et al., 2008) and having less than 5% of the overfitting items does

not affect item and person estimates substantially (Smith Jr., 2005), it should be reasonable to conclude that these items do not cause serious problems. These five items were not excluded from the GCT.

Table 10. Overfit items in the meaning section

Item No.	Difficulty (logits)	Model S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
10	-0.12	0.18	-2.8	0.77	-3.4	0.79
9	-0.40	0.18	-2.3	0.80	-2.4	0.86
36	-0.57	0.18	-2.2	0.82	-1.9	0.89
59	-0.78	0.18	-2.2	0.78	-2.7	0.84
8	-0.02	0.18	-1.9	0.83	-2.3	0.85

In summary, a total of eleven items were considered to be problematic and thus excluded from the GCT for future use of this test. This left a total of 49 acceptable items. The 49 acceptable items are broken down into 24 nouns, 13 verbs, 7 adjectives, and 5 adverbs with the approximate ratio of (noun): (verb): (adjective): (adverb) = 9:6:3:2. For each contextual clue, three or more items were acceptable. The subsequent section discusses the validity of the GCT.

4.6 Validity

This section aims to explain the validity of the GCT. Validity is generally viewed as a unitary concept that subsumes all validity under construct validity (APA, AERA, & NCME, 1999; Bachman, 1990; Chapelle, 1999; Messick, 1989, 1995). Messick (1989) states:

> Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment. (p.13)

Strictly speaking, it is not a test *per se* but rather the meaning of the test scores that is

validated. As Bachman (1990, p. 238) states, “in test validation we are not examining the validity of the test content or of even the test scores themselves, but rather the validity of the way we interpret or use the information gathered through the testing procedure.” In the present research, however, the phrase *validating a test* or *validity of a test* will be used instead of *validating the interpretation of test scores* or *validity of the interpretation of test scores* for convenience.

Construct validity as a unified concept may be addressed by means of providing evidence from various distinct aspects (e.g., Messick, 1989). In the present research, the GCT was validated based on Messick’s (1989, 1995) six aspects (content, substantive, structural, generalizability, external, and consequential) because his framework is increasingly being accepted as a useful basis for validation by researchers in language testing (Bachman, 1990, 2000; Bachman & Palmer, 1996; Chapelle, 1999; McNamara, 2006; Read & Chapelle, 2001) as well as in psychology and education (e.g., APA, AERA, & NCME, 1999). The two non-overlapping aspects (responsiveness and interpretability) proposed by the Medical Outcomes Trust Scientific Advisory Committee (1995) were also examined so that the issue of validity may be addressed in a more comprehensive way. Each of these eight aspects may in part be investigated effectively through Rasch measurement (Fisher Jr., 1994; Smith Jr., 2004b; Wolfe & Smith Jr., 2007). The subsequent sections attempt to provide evidence of the construct validity of the GCT from the eight aspects largely on the basis of Rasch measurement.

4.6.1 Content Aspect

The content aspect of construct validity aims to clarify “the boundaries of the construct domain to be assessed” (Messick, 1995, p. 745). This aspect addresses the relevance,

representativeness and technical quality of the items (Messick, 1989, 1995). Technical quality may be examined by Rasch item fit statistics (Smith Jr., 2004b). This was discussed in Section 4.5: the item fit analysis identified eleven misfit items which were excluded from the GCT. Thus, the remaining 49 items were considered to be acceptable in terms of item fit, which indicates a high degree of technical quality of the 49 acceptable items. Additional evidence may be provided by expert judgments (Wolfe & Smith Jr., 2007). This was examined by a series of pilot studies with a number of English teachers and PhD students in applied linguistics (see Section 3.7). The results in the pilot studies indicate that the items used in the GCT were the ones that most of the experts considered to be acceptable for a measure of the skill of guessing from context. The subsequent subsections discuss content relevance and representativeness.

Relevance

An in-depth discussion of the construct definition of guessing from context and the tasks for measuring the construct was given in the previous chapter. Here are the key points.

- The construct definition is based on strategies for guessing from context (Bruton & Samuda, 1981; Clarke & Nation, 1980; Williams, 1985): identifying the part of speech of the unknown word, looking for contextual clues, and deriving meaning.
- The three components (part of speech, contextual clue, and meaning) are teachable and available in every context.
- The GCT has three sections in order to measure each of the three components.
- The test words to be guessed were randomly selected from low-frequency words (words listed between the 11th and 14th 1,000 word families in the BNC word lists).
- The GCT focuses on four parts of speech (nouns, verbs, adjectives, and adverbs) because they account for the vast majority of English words.

- The ratio of the four parts of speech for the test words was (noun): (verb): (adjective): (adverb) = 9:6:3:2, in order to reflect actual language use.
- Contextual clues identified by previous studies may be categorised into twelve types. All of these clues were evenly included in the GCT.
- About half of the contextual clues appeared in the same sentence as the test word and the rest appeared outside of the sentence containing the test word.

It should be reasonable to conclude that the test content is highly relevant to the skill of guessing from context because the tasks were created so that guessing from context may be comprehensively measured.

Representativeness

Content-relevant tasks are not sufficient for valid measurement: the tasks need to be representative of the construct domain (Messick, 1995). The GCT is considered to be representative of the construct domain, because 1) the test words to be guessed were randomly selected from low-frequency words which were unlikely to be familiar to test-takers, 2) each test word was measured in a different passage so that a wide variety of words may be included (if test words were measured in a long passage, they would be semantically related under the same topic), 3) the ratio of the four parts of speech reflected actual language use, and 4) a wide variety of contextual clues were included.

Representativeness may be empirically evaluated by examining the Rasch item difficulty hierarchy (Smith Jr., 2004b). First, the spread of item calibrations was examined by item strata. An item strata index indicates the number of statistically different levels of item difficulty, and is derived by the following formula:

$$\text{Item strata} = (4\,G_{item}+1)/3,$$

where G_{item} is Rasch item separation. Item strata statistics need to be greater than 2.0

for useful tests, because "[i]f a sufficient (at least 2) number of item difficulty levels are unable to be identified, then one may have difficulty in interpreting the variable defined by the items" (Smith Jr., 2004b, p. 106). The item strata statistics for the three sections are presented in Table 11. This table shows that each section has more than two statistically distinct difficulty levels, which can be taken as supportive evidence for the representativeness of the tasks.

Table 11. Item strata for the three sections of the GCT

Section	Item strata
Part of speech	6.07
Contextual clue	3.57
Meaning	4.85

Another way of examining representativeness may be to see whether there are gaps in the item difficulty hierarchy. This may be addressed by looking at a person-item map (often called a Wright map in honour of a leading researcher in Rasch measurement, Ben Wright), which displays both persons (in terms of ability) and items (in terms of difficulty) on a Rasch interval scale.

Figure 11 is a person-item map for the part of speech section. The far left of this figure shows a Rasch logit (log odds of success) scale with the mean item difficulty being 0. This figure has two distributions on the logit scale: persons on the left and items on the right. More able persons and more difficult items are located towards the top and less able persons and less difficult items are located towards the bottom. For the person distribution, each # represents three persons and each * represents one or two persons. For the item distribution, each number indicates the unique item number for easy reference and the subsequent letter(s) indicates its part of speech (N=noun, V=verb,

	<More able persons>			<More difficult items>					
4	##########	T	+						
	####		|						
	*##		|						
			|						
			|						
	*####		|						
	##		|						
3	*#########		+	13-V					
	*##	S	|						
			|						
	*		|	T					
	*######		|						
	*########		|						
	*##		|	20-N					
2	*###		+	57-N					
	*#####		|	5-N	19-Adj				
	########		|	15-Adj	33-Adj				
	########		|	11-N					
	*##	M	|	32-N					
	*######		|	S					
	*###		|						
1	*#####		+	1-Adv					
	*########		|	24-N	58-N	59-Adv			
	######		|	45-Adj					
	*#		|	36-N					
	*####		|	18-Adv					
	*		|	17-Adj	28-N	48-Adv	51-N		
	*###	S	|	29-Adv					
0	###		+	M	40-N					
	*###		|	27-V					
	*#		|	26-N	30-Adj	39-N	46-V	50-V	56-N
	*#		|	2-Adj	35-N				
	#		|	31-V					
	#		|	6-V	25-V				
	*		|	8-V	38-N	42-V			
-1	*		+	14-Adj	23-N				
	*		|	7-N	37-N	44-N			
	*#	T	|	S	55-V				
	#		|	9-N					
	*		|	10-V					
	#		|	22-N					
	#		|	12-N					
-2	*		+						
	*		|	53-V					
	*		|	34-V					
	*		|						
	*		|	T					
			|						
			|						
-3	*		+						
	<Less able persons>			<Less difficult items>					

Note: N = noun, V = verb, Adj = adjective, Adv = adverb

Figure 11. Person-item map for the part of speech question

Adj=adjective, and Adv=adverb). For example, 13-V means that the unique item number is 13 and its part of speech is a verb. The two distributions are interrelated in that a person has a 50% probability of succeeding on an item located at the same point on the logit scale. This person's success probability increases for items located lower than that point, and vice versa. For example, a person with 0 logits has a 50% probability of succeeding on the item 40-N. This person has a greater success probability for items such as 27-V and 26-N which are located lower than 40-N. The M, S, and T in the middle represent the mean of the person or item estimates (M), one standard deviation from the mean (S), and two standard deviations from the mean (T). Figure 11 shows that there are few gaps in the item difficulty hierarchy lower than 2 logits, indicating a high degree of representativeness in terms of item difficulty for the range below 2 logits. The items do not adequately cover person abilities higher than 2 logits. This indicates a need for including more difficult items that are targeted to persons with high ability (more than 2 logits). However, the test was created so that difficult items would also be included. Here are the three most difficult items for the part of speech section.

Item 13 (Test word: *vanink*; Original word: *abound*).
The fact that birds **vanink** means that the woods are a good place to discover various kinds of birds.
(1) noun (2) verb (3) adjective (4) adverb

Item 20 (Test word: *tarrand*; Original word: *amnesia*).
"Too much work will make you very tired, and suddenly cause **tarrand**: you won't remember anything and you even forget who you are.
(1) noun (2) verb (3) adjective (4) adverb

Item 57 (Test word: *duterages*; Original word: *beverages*).
From the 10th to 25th of October the show is held about various ways of having **duterages** such as tea and coffee.
(1) noun (2) verb (3) adjective (4) adverb

These items may be difficult because of their grammatical complexity. For Item 13, the test word is embedded in a subordinate clause. For Item 20, the test word is the object of the verb *cause* whose subject is located eight words away from it. For Item 57, the test word is the object of the gerund *having* which might be mistaken as an indication of a perfect tense. This indicates that difficult items are included in the test, and if a person gets these difficult items correct, then he or she may be regarded as having sufficient knowledge for identifying the part of speech of a word in a sentence. Together with the large item strata (6.07), the part of speech section may be acceptably representative of the construct being mesured.

Figure 12 is a person-item map for the contextual clue section. This figure shows that there are few gaps in the item difficulty hierarchy and the items largely cover the range of person abilities. Although the spread of item difficulties is not as wide as that of the part of speech section, this may not cause a problem because of the acceptable item strata (3.57). This may be taken as supportive evidence for the representativeness of the items in the contextual clue section.

Figure 13 is a person-item map for the meaning section. As with the contextual question, this figure shows that there are few gaps in the item difficulty hierarchy for the meaning section and the items largely cover the range of person abilities. This may be taken as supportive evidence for the representativeness of the items in the meaning section.

This subsection has looked at the content aspect of construct validity. Logical and empirical evidence indicates that the GCT is relevant to and representative of the construct of the skill of guessing from context.

<More able persons>			<More difficult items>					
3	*	+						
		|						
		|						
		|						
	*	|						
	#	|						
		|						
	*#	|						
2	*#	T +						
	##	|						
	*	|						
	#	|	5-CC-O					
	*#	|						
		|						
	*####	|	30-CE-O					
	##	S |	T 24-ID-O	28-CE-I				
1	*######	+						
	*#	|	29-CE-O					
	*******	|	14-DD-O					
	#	|						
	*******	|	S 34-RE-O					
	*####	|	2-CC-I					
	########	M |	1-CC-I	22-ID-I	39-MO-I	45-RF-O		
	####	|	7-SY-I	11-DD-I	33-RE-O	46-WS-I	51-AS-I	
0	*****	+	M 6-SY-I	18-AP-I	35-RE-O	37-MO-I	40-MO-I	
		|	8-SY-I	10-SY-O	13-DD-I	15-DD-O		
	******	|	17-AP-I	20-AP-I	50-WS-I	59-EX-O		
	*##	|	23-ID-I	32-RE-O	38-MO-I	42-RF-I	56-EX-I	58-EX-I
	##########	|	55-AS-I					
	*#	|	S 9-SY-O	12-DD-I	25-ID-O	31-RE-O	53-AS-I	
	******	S |	19-AP-I					
	*#	|	27-CE-I					
	*##	|	26-CE-I	36-MO-I				
-1	*##	+	44-RF-O	48-WS-I				
	*##	|	T 57-EX-I					
	*#	|						
	*	|						
	*	|						
	#	T |						
	#	|						
		|						
-2		+						
	*	|						
<Less able persons>			<Less difficult items>					

Note: AP = Appositive, AS = Association, CC = Contrast/comparison, CE = Cause/effect, DD = Direct description, EX = Example, ID = Indirect description, MO = Modification, RE = Restatement, RF = Reference, SY = Synonym, WS = Words in series, I = Inside (the contextual clue appeared in the same sentence as the test word), O = Outside (the contextual clue appeared outside of the sentence containing the test word).

Figure 12. Person-item map for the clue section

<More able persons>			<More difficult items>			
*		|				
3		+				
*		|				
		|				
		|				
*		|				
		|				
		|				
*		|				
2	#	+				
#		|				
		|				
*#	T	|	40-N-MO-I			
*		|				
#		|	T 34-V-RE-O			
*##		|	14-Adj-DD-O			
*#		|	28-N-CE-I	30-Adj-CE-O		
1	*###	+	7-N-SY-I	39-N-MO-I	42-V-RF-I	
*#		|				
#####	S	|	13-V-DD-I			
*#		|	S			
*######		|	24-N-ID-O			
*#		|	1-Adv-CC-I	11-N-DD-I	26-N-CE-I	32-N-RE-O
*########		|	2-N-CC-I	15-Adj-DD-O	33-Adj-RE-O	38-N-MO-I
*######		|	17-Adj-AP-I	53-V-AS-I		
0	####	+	M 5-N-CC-O	8-V-SY-I	58-N-EX-I	
	##########	M |	10-V-SY-O	22-N-ID-I	23-N-ID-I	
		|	25-V-ID-O	50-V-WS-I	56-N-EX-I	
##		|	18-Adv-AP-I	31-V-RE-O	35-N-RE-O	37-N-MO-I
*##########		|	6-V-SY-I	9-N-SY-O	29-Adv-CE-O	
*####		|	19-Adj-AP-I	46-V-WS-I	51-N-AS-I	
*#####		|	S 20-N-AP-I	36-N-MO-I	45-Adj-RF-O	57-N-EX-I
######		|	55-V-AS-I	59-Adv-EX-O		
*###		|	27-V-CE-I			
-1	*###	S +	12-N-DD-I	48-Adv-WS-I		
*####		|				
*##		|	44-N-RF-O			
*#		|	T			
#		|				
		|				
*#		|				
#	T	|				
-2	*	+				
*		|				
<Less able persons>			<Less difficult items>			

Note: N = noun, V = verb, Adj = adjective, Adv = adverb, AP = Appositive, AS = Association, CC = Contrast/comparison, CE = Cause/effect, DD = Direct description, EX = Example, ID = Indirect description, MO = Modification, RE = Restatement, RF = Reference, SY = Synonym, WS = Words in series, I = Inside (the contextual clue appeared in the same sentence as the test word), O = Outside (the contextual clue appeared outside of the sentence containing the test word)

Figure 13. Person-item map for the meaning section

4.6.2 Substantive Aspect

The substantive aspect of construct validity refers to "theoretical rationales for the observed consistencies in test responses [...] along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks" (Messick, 1995, p. 745). This aspect may be evaluated by examining whether the empirical item hierarchy is presented as predicted by theoretical argument and whether each person's response pattern is consistent with that item hierarchy (Smith Jr., 2004b). To begin with, the relationship between theoretical and empirical item hierarchy was examined for each of the three sections.

For the part of speech section, it was hypothesised that adjectives and adverbs would be more difficult to identify than nouns and verbs. Liu and Nation (1985) argue that nouns and verbs may be easier to guess than adjectives and adverbs, because nouns and verbs have wider ranging relationships with other parts of the context than adjectives and adverbs. Aborn, Rubenstein, and Sterling (1959) found that their participants guessed the meanings of unknown words with wrong parts of speech more frequently in the case of adjectives and adverbs than nouns and verbs. The hypothesis about the difficulty order of the four parts of speech was tested by comparing the mean Rasch item difficulty estimates for the four parts of speech. Figure 14 shows the mean item difficulty with 95% confidence intervals for the four parts of speech. Larger numbers in logits on the vertical axis indicate more difficult items, and vice versa. This figure shows that the mean item difficulties were higher for adjectives and adverbs than nouns and verbs; however, a one-way ANOVA did not detect a statistically significant difference between the mean item difficulties of the four parts of speech ($F(3,44) = 2.514$, $p = .070$). This may be because other factors such as grammatical complexity of

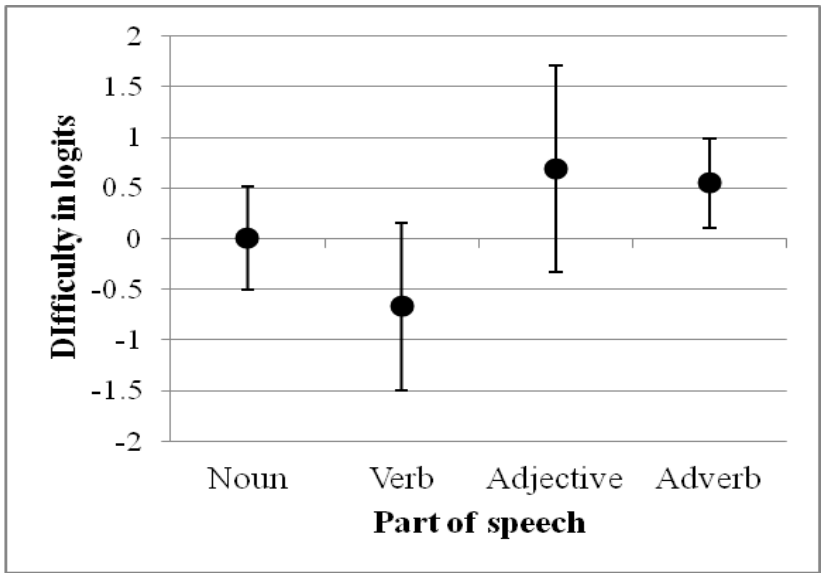


Figure 14. Mean difficulties and 95% confidence intervals of the part of speech question according to part of speech

a sentence are involved in identifying the part of speech of a word.

Another factor that may affect the item difficulty is the presence or the absence of derivational and inflectional suffixes. Research (de Bot, et al., 1997; Paribakht & Wesche, 1999) has indicated that learners typically integrate information from the word parts and the context in guessing the meaning of an unknown word. The GCT used nonsense words with real suffixes for the test words in order to avoid ambiguity of correct answers and to increase ecological validity (see Section 3.5.1). Thus, it was hypothesised that the part of speech of suffixed words would be answered more correctly than that of non-suffixed words. A *t*-test (2-tailed) was performed to test the hypothesis (Table 12). The results showed that suffixed-word items were easier (-0.29 logits) than non-suffixed-word items (0.16 logits), but no significant difference was found between them. Taken together, the two hypotheses about the difficulty order of items in the part of speech section were not rejected by the empirical evidence, but were not supported statistically.

Table 12. Difference between items of suffixed and non-suffixed words

	N	*Mean*	*S.D.*	*t*	*d.f.*	*p*
Suffixed	20	-0.29	1.13	-1.24	47	.222
Non-suffixed	29	0.16	1.31			

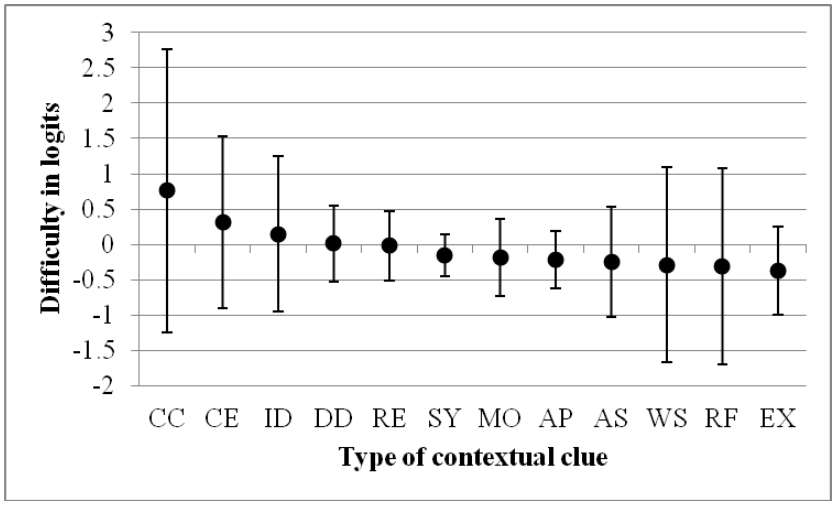
For the contextual clue section, it was hypothesised that contextual clues appearing in the same sentence as the test word would be easier to find than those appearing outside of the sentence containing the test word. Carnine, Kameenui, and Coyle (1984) demonstrated that the closer a contextual clue was to the unknown word, the more likely it was that learners were successful in guessing. A *t*-test (2-tailed) was performed to test the hypothesis (Table 13). The results showed that the clue-inside (clues in the same sentence as the test word) items were significantly easier to find than the clue-outside (clues outside of the sentence containing the test word) items ($\alpha = .05$). This may be taken as supportive evidence for the substantive aspect of construct validity of the contextual clue section.

Table 13. Difference between clue-inside and clue-outside items

	N	*Mean*	*S.D.*	*t*	*d.f.*	*p*
Clue-inside	32	-0.17	0.42	-2.19	47	.033
Clue-outside	17	0.19	0.71			

Another factor that may affect the success of the tasks in the contextual clue section is the explicitness of the contextual clue: more explicit clues may be less difficult, and vice versa. Carnine, Kameenui, and Coyle (1984) compared the number of successful guesses based on explicit clues (synonym and contrast clues) and implicit clues (inference clues). They found that unknown words with synonym clues (clues derived from a word or phrase that has essentially the same meaning as the unknown word) were significantly easier to guess than unknown words with inference clues (clues derived by deduction from information in the context). However, they did not

find a significant difference between contrast clues (clues derived from a word or phrase that has essentially the opposite meaning to the unknown word) and the other two clues. Their findings indicate that the explicitness of the contextual clue may not be highly predictive of the difficulty order of the items in the contextual clue section. Thus, it was hypothesised that there would be no clear tendency that more explicit contextual clues would be easier to find than less explicit ones. Figure 15 shows the mean item difficulty estimates for the twelve contextual clues with 95% confidence intervals. The ID (indirect description) and the RE (restatement) clues are considered to be less explicit than the others, because these two clues have no explicit signals that indicate the relationships with other parts of the context. This figure shows that there was little difference in difficulty between the less explicit clues (ID and RE) and the others. A one-way ANOVA did not detect a statistically significant difference between any two mean item difficulties of the twelve contextual clues ($F(11,37) = 1.200$, $p = .321$). This



Note: AP = Appositive, AS = Association, CC = Contrast/comparison, CE = Cause/effect, DD = Direct description, EX = Example, ID = Indirect description, MO = Modification, RE = Restatement, RF = Reference, SY = Synonym, WS = Words in series.

Figure 15. Mean difficulties and 95% confidence intervals of the contextual clue question according to the type of contextual clue

may be taken as supportive evidence for the hypothesis.

For the meaning section, it was difficult to predict item difficulty because guessing the meaning of unknown words is affected by various factors such as grammatical knowledge and discourse knowledge (de Bot, et al., 1997; Haastrup, 1987, 1991; Nassaji, 2003). It was hypothesised that no single factor would be sufficient for determining the difficulty order of the items in the meaning section. One important factor is grammatical knowledge, or knowledge of part of speech. The part of speech of the unknown word may affect guessing, but as shown in Table 14 previous studies are not consistent as to the difficulty order according to part of speech. Based on the previous studies, however, it was predicted that the meaning of adjectives would be more difficult to guess than that of adverbs, because all the previous studies indicate that adjectives are more difficult than adverbs. Figure 16 shows the mean item difficulty in logits with 95% confidence intervals accrording to part of speech. This figure shows that the meaning of adjectives was more difficult to guess than that of adverbs, but a one-way ANOVA did not detect a significant difference between them ($F(3) = 0.992$, $p = .405$). This may partly support the hypothesis that the part of speech alone cannot determine the difficulty order of the items in the meaning section.

Table 14. Difficulty order of guessing the meaning of unknown words according to part of speech

	Aborn, Rubenstein, & Sterling (1959)	Dulin (1969)	Liu & Nation (1985)
Hardest	Adjective	Verb	Adjective
	Noun	Adjective	Adverb
	Adverb	Adverb	Noun
Easiest	Verb	Noun	Verb

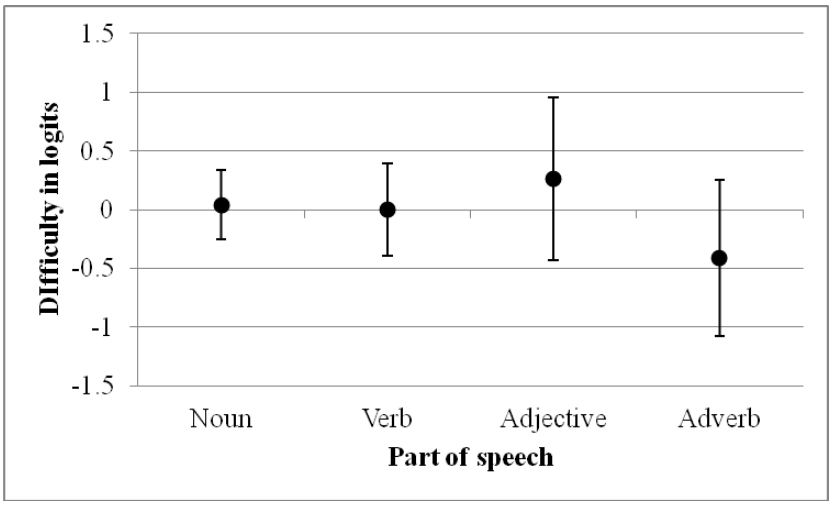


Figure 16. Mean difficulties and 95% confidence intervals of the meaning section according to part of speech

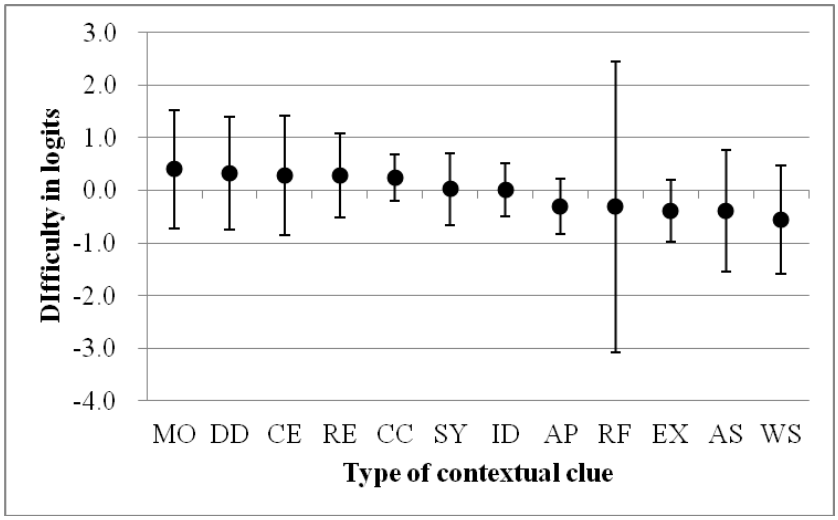
The difficulty order of the items in the meaning section may also be affected by the place of a contextual clue; that is, the meaning of an unknown word may be easier to guess when the contextual clue appears closer to the unknown word (Carnine, et al., 1984). It was predicted that there would be a tendency that the meaning of an unknown word is easier to guess when the contextual clue is closer to the unknown word. A *t*-test (2-tailed) was performed to test the hypothesis (Table 15). The results showed that the clue-inside (clues in the same sentence as the test word) items (0.00 logits) were slightly easier to guess than the clue-outside (clues outside of the sentence containing the test word) items (0.05 logits), but no significant difference was found between them. The prediction may not be fully supported by the empirical evidence, but this may support the hypothesis that no single factor would be able to determine the difficulty order of the items in the meaning section.

Table 15. Difference between clue-inside and clue-outside items

	N	*Mean*	*S.D.*	*t*	*d.f*	*p*
Clue-inside	32	0.00	0.66	-0.243	47	.809
Clue-outside	17	0.05	0.71			

Finally, the difficulty order of the items in the meaning section was examined to see if it was affected by the explicitness of a contextual clue (Carnine, et al., 1984). It was predicted that less explicit clues would make guessing more difficult but the tendency would be weak. Figure 17 shows that the mean difficulty estimates of the items with less explicit clues (RE and ID) were slightly higher than 0 logits (the average difficulty of all items), but these items were not typically difficult. Taken together, the empirical evidence may be taken as supporting the hypothesis that no single factor is sufficient for determining the difficulty order of the items in the meaning section.

Another way of evaluating the substantive aspect of construct validity of the items in the meaning section is to examine the degree to which grammatical and discourse knowledge contributes to performance on the items. It was hypothesised that the performance on the meaning items would have a closer relationship with the performance on the contextual clue items than the performance on the part of speech



Note: AP = Appositive, AS = Association, CC = Contrast/comparison, CE = Cause/effect, DD = Direct description, EX = Example, ID = Indirect description, MO = Modification, RE = Restatement, RF = Reference, SY = Synonym, WS = Words in series.

Figure 17. Mean difficulties and 95% confidence intervals of the meaning question according to the type of contextual clue

items. Being able to identify the part of speech of the unknown word allows a 'Who does what to whom' analysis, but this may only be helpful for deriving a partial meaning such as positive/negative or person/thing, rather than a precise meaning (Clarke & Nation, 1980). In many cases, contextual clues are necessary for deriving a precise meaning. This hypothesis was examined using a multiple regression analysis, where the dependent variable was the person ability estimates from the meaning section and the independent variables were the person ability estimates from the part of speech section and from the contextual clue section. Figure 18 presents a path diagram of the multiple regression analysis (without correction for attenuation due to measurement error).[10] This figure shows that the β coefficient for the contextual clue section (.44) was higher than that for the part of speech section (.32), which may be taken as supportive evidence for the hypothesis.

The substantive aspect of construct validity was also evaluated by examining the consistency of each person's response pattern with the item hierarchy. More specifically,

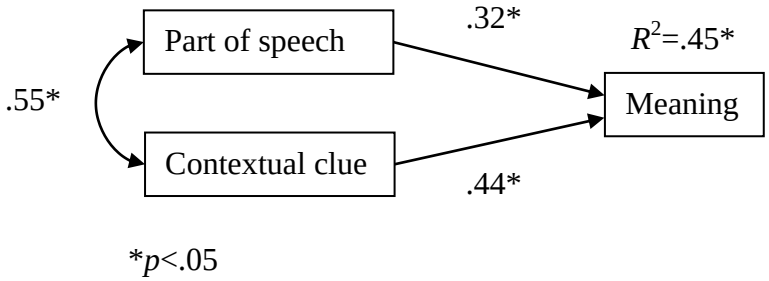


Figure 18. Relationships of the part of speech and the contextual clue sections to the meaning section

[10] No serious sign of multi-collinearity was detected. The variance inflation factor (VIF) was 1.45 for both the part of speech and the contextual clue sections, which is well below 10 which is generally taken as the threshold for multi-collinearity.

Rasch person fit statistics were calculated for each section. Person fit examines the degree of match between the observed responses and the theoretical model that requires a person of a given ability to have a greater probability of a higher rating on easier items than on more difficult items (Smith Jr., 2004b). As with item fit, a misfit person was defined as outfit $t > 2.0$ or infit $t > 2.0$ (underfit), or outfit $t < -2.0$ or infit $t < -2.0$ (overfit). Table 16 presents the number of misfit persons for each section. Each section had the misfit rate of less than 5% which was expected to occur by chance given the nature of the *z* distribution. This indicates that the test-takers' response pattern corresponded to the modelled difficulty order, which may be taken as supportive evidence for substantive aspect of construct validity.[11]

Table 16. Number of misfit persons

Section	Number of underfit persons	Number of overfit persons	Total	%
Part of speech	13	2	15	3.5
Contextual clue	5	3	8	1.9
Meaning	11	4	15	3.5

4.6.3 Structural Aspect

The structural aspect of construct validity "appraises the fidelity of the scoring structure to the structure of the construct domain at issue" (Messick, 1995, p. 745). The evaluation of this aspect may be addressed by examining the unidimensionality (the degree to which a test measures one attribute at a time) of the intended structure, because a unidimensional measure allows a straightforward scoring method; that is, the

[11] Future research could be carried out to collect further evidence for the substantive aspect by conducting qualitative research including think-aloud protocols which investigates the degree to which the theoretical processes reflect respondents' actual processes.

cumulative total raw scores obtained simply by counting the observed responses are sufficient for estimating item difficulty and person ability (Smith Jr., 2004b; Wolfe & Smith Jr., 2007). Several studies (Slinde & Linn, 1979; Smith Jr., 2004a; Smith & Miao, 1994) have argued that Rasch analysis is superior to factor analytic methods in assessing dimensionality, because unlike factor analytic methods, Rasch models do not assume a normal distribution of the data or multiple parameters (e.g., item discrimination and guessing). For this reason, dimensionality was assessed based on Rasch analysis.

As there is no agreed-upon method for assessing dimensionality, dimensionality was examined from a number of perspectives. Linacre (1995) suggested that dimensionality may be addressed by 1) item correlations, 2) fit statistics, and 3) principal components analysis (PCA) of standardised residuals without rotation. An item correlation examines the degree to which the items are aligned in the same direction on the latent variable. This was investigated by computing the point-measure correlation (correlation between the observations on an item and the corresponding person ability estimates). No items showed unacceptably low point-measure correlations (<.10).[12] This indicates that unidimensionality holds in terms of item correlation.

A second way of investigating dimensionality is to identify misfit items based on fit statistics. As mentioned in Section 4.5, items with outfit $t > 2.0$ or infit $t > 2.0$ were identified as misfitting to the Rasch model and were excluded from the GCT. Thus, the remaining items may conform to the model which requires that measures be unidimensional. Item fit analysis may be the most reliable of the three approaches to

[12] One item in the meaning section had a point-measure correlation of .09 but this item was excluded from the GCT (see Section 4.5.3).

detecting dimensionality. Research based on simulated data (Smith, 1996; Smith & Miao, 1994) has shown that the Rasch item fit approach detects dimensionality more accurately than other approaches including the PCA when the intention is to create a unidimensional variable where few items are expected to contribute to the second component and the correlation between the components are expected to be high. This may be taken as strong evidence for unidimensionality.

Finally, the PCA of standardised residuals was performed for each section in order to examine whether there was only a small amount of variance in the residuals accounted for by other components (dimensions) than the Rasch model which extracts the first major component in the observations.[13] An acceptable first contrast (largest secondary component) should have an eigenvalue (standardised residual variance) of less than 3 which means that less than three items are loading onto another dimension (Linacre, 2010a, p. 444; Linacre & Tennant, 2009). A simulation study (Raîche, 2005) demonstrates that an eigenvalue of 2 is possible at the random level. The eigenvalue of the first contrast was 2.0 for the part of speech section, 2.0 for the contextual clue section, and 1.9 for the meaning section, indicating that the scores generated by the three sections are acceptably unidimensional. In addition, dimensionality may be assessed by examining whether the eigenvalues of other contrasts reach an asymptote at the first contrast (Stevens, 2002; Wolfe & Smith Jr., 2007). Figures 19-21 show the scree plot for each section. All three figures indicate that when the first largest component (the Rasch component) was extracted, the eigenvalues of the other components (the 1st to the 5th contrasts) reached an asymptote at the first contrast. Taken together, the scores generated by the GCT indicate a high degree of

[13] Unidimensionality depends on the size of the second component, and not on the magnitude of the variance explained by the first (Rasch) component, because unidimensionality is not degraded by the unexplained variance if most of it is random noise (Linacre, 2010a, p.440).

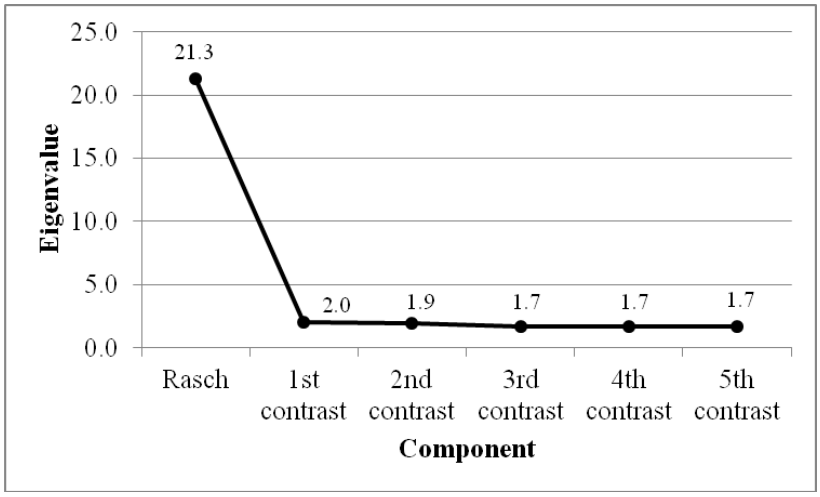


Figure 19. Scree plot for the part of speech section

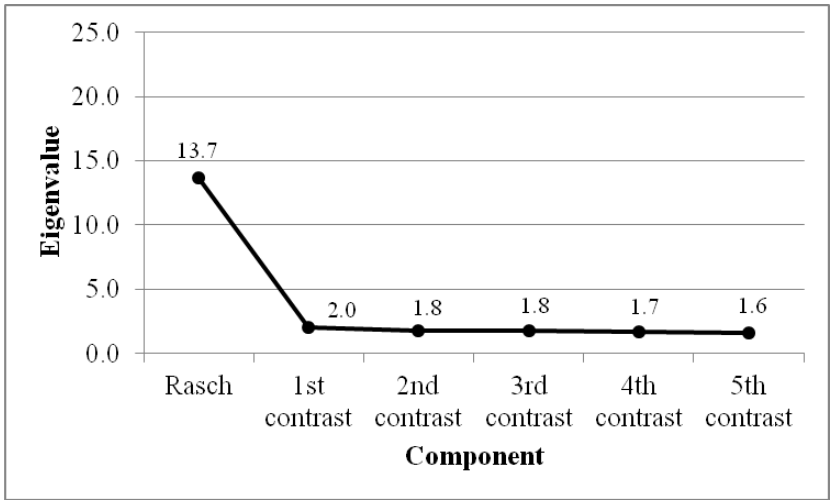


Figure 20. Scree plot for the contextual clue section

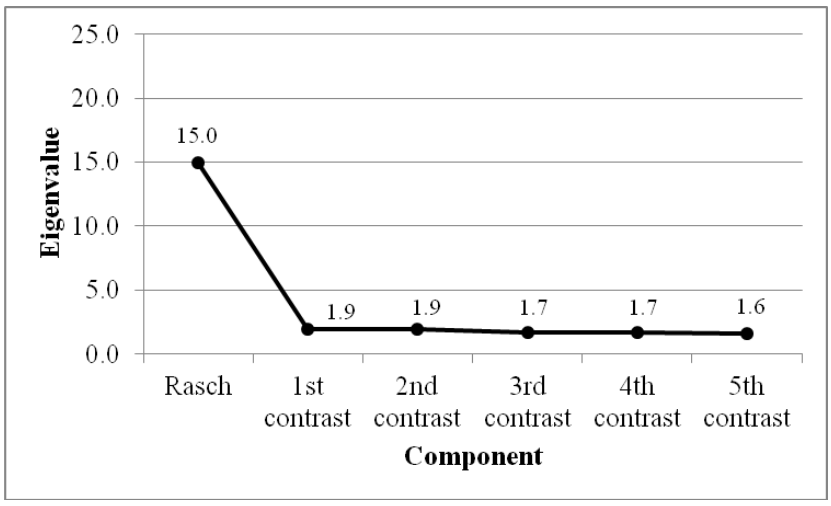


Figure 21. Scree plot for the meaning section

unidimensionality. This may serve as positive evidence for the structural aspect of construct validity of the GCT.

4.6.4 Generalizability Aspect

The generalizability aspect of construct validity deals with "the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks" (Messick, 1995, p. 745). In Rasch measurement, this aspect may be approached by examining the extent to which item difficulty and person ability estimates are invariant within the measurement error across measurement contexts such as different groups of examinees, time, or tasks (Andrich, 1988; Smith Jr., 2004b; Wolfe & Smith Jr., 2007; Wright & Stone, 1979). Wolfe and Smith Jr. (2007) divided this aspect into four subcategories: item calibration invariance (stability of item difficulty estimates), person measure invariance (stability of person ability estimates), reliability, and invariance across administrative contexts. Each of these subcategories will be examined in the following subsections.

Item Calibration Invariance

The invariance of item calibrations refers to "the degree to which item calibrations maintain their meaning and interpretability [...] across groups of respondents and across time" (Wolfe & Smith Jr., 2007, p. 215). This was investigated by analysing uniform differential item functioning (DIF), an indication of unexpected behaviour by items showing that item calibrations vary across samples by more than the modelled error (Bond & Fox, 2007; Linacre, 2010a; Wolfe & Smith Jr., 2007). The DIF analysis was performed through a *t*-test approach instead of a Mantel-Haenszel approach (Mantel,

1963; Mantel & Haenszel, 1959), because the data were not complete in that the design allowed systematic missing data (Linacre, 2010a, p. 490).

First, the DIF analysis was performed in order to examine whether the item calibrations from male (N = 277) and female (N = 151) test-takers varied widely for each of the three sections. Welch's t-test revealed that statistically significant DIF was detected for one item in each section. Table 17 presents the Rasch difficulty estimates and the Welch's t statistics for the three items with significant DIF ($\alpha = .05$).

Table 17. DIF analysis for gender

Section	Item No.	Difficulty estimates for males (logits)	Difficulty estimates for females (logits)	*t*	*d.f.*	*p*
Part of speech	24	0.34	1.46	-2.36	105	.020
Contextual clue	58	-0.44	0.46	-2.09	76	.040
Meaning	30	0.81	1.70	-2.00	115	.048

Here are the three items that displayed significant DIF.

Item 24 (Test word: *vansel*; Original word: *tremor*).

She smelled something awful from his body, and a **vansel** ran through her.
(1) noun (2) verb (3) adjective (4) adverb

Item 58 (Test word: *delincert*; Original word: *anaesthesia*).

About 30 years ago, people were patient about everything and did not (1)complain about anything. When my mother was a (2)child in the 1930's, doctors did not like to use **delincert**; for example, the doctor took out her first set of teeth without using it even if it was very (3)painful.
(1) complain
(2) child
(3) painful

Item 30 (Test word: *strocastic*; Original word: *ophthalmic*).

> Teachers may want to give as much help as possible to students who have difficulty seeing things, but there has been little information about the problems these students can face. For this reason, some basic **strocastic** information is given that will help teachers to offer effective learning materials for them.
> (1) relating to education
> (2) relating to eyes
> (3) relating to computers

For all three items, males achieved significantly higher success probabilities (lower difficulty estimates) than females. There seems to be no clear reason for this item bias. It should be noted that the overall number of DIF items is not statistically problematic, because each section has only one DIF item out of the 49 acceptable items (2.0%), which is less than 5% which may occur by chance given the nature of Type I error.

DIF was also investigated in terms of test-takers' native language; that is, whether the item calibrations from Japanese learners and learners with different L1 background varied widely for each of the three sections. A total of 30 participants with other native languages than Japanese took one of the six 20-item forms. They were international students at Japanese universities from other countries. Their native languages included Chinese (23), Korean (4), French (2), and Spanish (1). No DIF was found for any items in the three sections. This may be taken as supportive evidence for item calibration invariance across groups of different native languages; however, the results need to be interpreted carefully. The reference group (the group of participants with different native languages from Japanese) was biased towards Chinese people and might not be generalizable to other groups of people. In addition, the small size of the reference group may have affected the DIF analysis. A simulation study indicates that DIF analyses require more than 200 respondents per group for obtaining adequate (>80% power) performance (Scott et al., 2009). In fact, eleven items had a difference in item

calibration between the two groups (Japanese vs. other) by more than 1.0 logit for the part of speech section, five items for the contextual clue section, and three items for the meaning section. No statistically significant difference was found for these items perhaps due to large standard errors with a small number of people for the reference group. Further evidence needs to be collected for this aspect.

Person Measure Invariance

The invariance of person ability estimates was examined by analysing differential person functioning (DPF), an indication of unexpected behaviour by persons showing that person measures vary across different situations by more than the modelled error (Bond & Fox, 2007; Linacre, 2010a). Specifically, the items were divided into two halves (the first half and the second half) in order to examine whether person ability estimates were affected by the effects of practice or fatigue. As with DIF, the DPF analysis was performed through a *t*-test approach instead of a Mantel-Haenszel approach. The results showed that no statistically significant DPF was detected for any persons for the three sections ($\alpha = .05$). In other words, no practice or fatigue effect was observed statistically. This may be taken as supportive evidence for person measure invariance; however, the results need to be interpreted carefully. The person abilities were estimated based on six to ten items for each group after the deletion of the misfit items. This small number of items for person ability estimation may have caused large standard errors, which might have been the reason for being unable to detect any significant DPF.

Reliability

A third way of investigating the generalizability aspect of construct validity is to examine the degree of reliability; that is, "reproducibility of relative measure location" (Linacre, 2010a, p. 511). Rasch person reliability, which is equivalent to traditional reliability coefficients such as Cronbach's alpha, KR-20, and the Generalizability coefficient, was computed for each of the six forms in the three sections. Rasch person separation was also calculated because it is linear and ranges from zero to infinite. The conventional reliability estimates are non-linear and suffer from ceiling effects within the range between zero and one (Smith Jr., 2004b). Tables 18-20 present the Rasch person separation and reliability estimates for each of the six forms (Forms A to F) in each of the three sections after the deletion of the misfit items.

Table 18. Rasch person separation and reliability for the part of speech section

	No. of items	No. of participants	Person separation	Person reliability
Form A	17	71	1.67	.74
Form B	19	68	1.47	.68
Form C	13	76	1.12	.56
Form D	15	76	1.87	.78
Form E	18	57	1.39	.66
Form F	16	80	1.23	.60

Table 19. Rasch person separation and reliability for the contextual clue section

	No. of items	No. of participants	Person separation	Person reliability
Form A	17	71	1.21	.59
Form B	19	68	1.47	.68
Form C	13	76	1.25	.61
Form D	15	76	1.58	.71
Form E	18	57	1.39	.66
Form F	16	80	1.47	.68

Table 20. Rasch person separation and reliability for the meaning section

	No. of items	No. of participants	Person separation	Person reliability
Form A	17	71	1.44	.67
Form B	19	68	1.24	.60
Form C	13	76	1.32	.63
Form D	15	76	1.41	.67
Form E	18	57	1.79	.76
Form F	16	80	1.39	.66

The results showed that the reliability estimates ranged between .56 and .78 with the average being .66. This low reliability may have been caused by a small number of items (Linacre, 2010a). Each form had 20 items from which misfit items were excluded for reliability estimates, because pilot studies indicated that the results might be less reliable due to fatigue effects with more than 20 items. For future use of the GCT, the test length needs to be determined in order to achieve a certain degree of reliability (see Section 4.7 for further discussion).

The average reliability of .66 may not be unacceptably low. Fukkink and de Glopper (1998) conducted a meta-analysis of twelve previous studies on the effects of teaching on guessing from context, and reported that the tests used in these studies had the average reliability estimate of Cronbach's alpha = .63 (Max=.85, Min=.13). The low reliability estimates may be understandable, because the construct of guessing from context involves a wide range of language ability including reading skills and knowledge of vocabulary and grammar. Bachman and Palmer (1996) argue:

> "If the construct definition focuses on a relatively narrow range of components of language ability, the test developer can reasonably expect to achieve higher levels of reliability than if the construct is complex, including a wide range of components of language ability, as well as topical knowledge." (p.135)

This would suggest that a reliability of .66 may be reasonable for the GCT. Taken together, the reliability estimates may be acceptable given the nature of the test

(complex construct) and higher reliability may be expected by increasing the number of items per test form for future use of the GCT.

Reliability was also examined for item calibrations. Rasch item reliability, which has no traditional equivalent, addresses the degree to which item difficulties are reproducible. High item reliability is caused by a large sample and a wide variety of item difficulties (Linacre, 2010a). Tables 21-23 present the Rasch item separation and reliability estimates for each of the six forms (Forms A to F) in each of the three sections after the deletion of the misfit items.

Table 21. Rasch item separation and reliability for the part of speech section

	No. of items	No. of participants	Item separation	Item reliability
Form A	17	71	3.80	.94
Form B	19	68	2.71	.88
Form C	13	76	2.90	.89
Form D	15	76	3.65	.93
Form E	18	57	2.74	.88
Form F	16	80	2.78	.89

Table 22. Rasch item separation and reliability for the contextual clue section

	No. of items	No. of participants	Item separation	Item reliability
Form A	17	71	1.50	.69
Form B	19	68	2.17	.82
Form C	13	76	1.77	.76
Form D	15	76	2.19	.83
Form E	18	57	2.20	.83
Form F	16	80	2.71	.88

Table 23. Rasch item separation and reliability for the meaning section

	No. of items	No. of participants	Item separation	Item reliability
Form A	17	71	1.83	.77
Form B	19	68	2.37	.85
Form C	13	76	2.18	.83
Form D	15	76	1.80	.76
Form E	18	57	2.51	.86
Form F	16	80	3.14	.91

The results showed that the reliability estimates ranged between .69 and .94 with the average being .84. This indicates that the item difficulty estimates are reproducible to a high degree.

Invariance across Administrative Contexts

A final way of evaluating the generalizability aspect is to examine the stability of performance across administrative contexts. This may be approached "by comparing average measures observed between contextual situations or decision-making contexts using hypothesis testing" (Wolfe & Smith Jr., 2007, p. 217). For future use of the GCT, person ability may need to be estimated based on the performance on the items that a learner actually took, without using intentional missing data as designed for the present research. Thus, administrative invariance was evaluated by examining the degree to which the person ability estimates from the short version (the items that the test-takers actually took) were consistent with those from the long version (the overall 49 acceptable items based on the missing data design). A paired *t*-test was performed for each section in order to investigate whether statistically significant difference was found between the person ability estimates from these two versions. Table 24 presents the mean Rasch person ability estimates in logits for the two versions, *t*-statistics, and the point-biserial correlation coefficient *r* for an indication of effect size as calculated by the following formula:

$$r = \sqrt{\frac{t^2}{t^2 + df}},$$

where t = *t*-statistic in the *t*-test and *d.f.* = degree of freedom.

Table 24. Rasch person measures, *t*-statistics, and effect size between the short and long versions for the three sections

	Short version		Long version		*t*	*d.f.*	*p*	*r*
	M	*S.D.*	*M*	*S.D.*				
Part of speech	1.83	1.82	1.70	1.65	4.62	427	.000	.218
Contextual clue	0.36	1.15	0.28	1.01	4.77	427	.000	.225
Meaning	-0.07	1.08	-0.07	0.91	0.00	427	.998	.001

N=428.

Table 24 shows that the person ability was estimated significantly higher with the short version than with the long version for the part of speech and the contextual clue sections. This statistical significance may have been detected due to a large sample size (*N*=428). The difference between the mean person abilities estimated from the two versions was small: 0.13 logits for the part of speech section and 0.08 logits for the contextual clue section. In addition, the effect sizes (.218 and .225) were found to be small (Cohen, 1988, 1992). This indicates that the difference between the short and the long versions for the part of speech and the contextual clue sections may be acceptably small so that the short versions would produce person ability estimates approximate to the long version. For the meaning section, no significant difference was found between the short and the long versions ($\alpha = .05$), and person ability estimates from the short version was much the same as those from the long version ($r = .001$). Taken together, the short version of the GCT produced person ability estimates approximate to the long version with the missing data design, although the short versions slightly overestimated person abilities for the part of speech and contextual clue sections. This may be taken as supportive evidence for invariance across administrative contexts.[14]

[14] Evidence may also be collected by examining whether person ability estimates are stable between paper- and computer-based formats (Wolfe & Smith Jr., 2007). In the present research, a paper-based format was used because the test was administered in intact English classes where computer facilities were not available in many classrooms. Future research may include data from a computer-based format to add empirical evidence for construct validity for this aspect.

This subsection has looked at the generalizability aspect of construct validity. Although further evidence may need to be collected for L1-related DIF and the effects of practice or fatigue, the empirical evidence from this aspect largely supports the validity of the GCT.

4.6.5 External Aspect

The external aspect refers to "the extent to which the test's relationships with other tests and nontest behaviors reflect the expected high, low, and interactive relations implied in the theory of the construct being assessed" (Messick, 1989, p. 45). This may be addressed by examining convergent and discriminant correlations with external variables (Messick, 1995). Convergent evidence is derived by showing the correspondence between different measures of the same construct. Discriminant evidence, on the other hand, is derived by showing the lack of correspondence from measures of distinct constructs.

First, in an attempt to provide convergent evidence, the correlation between the receptive and the productive versions of the GCT was examined. The receptive version refers to the original GCT in multiple-choice format. The productive version refers to the modified GCT where all the options were omitted and test-takers had to write answers instead of choosing answers. It was hypothesised that the scores from these two versions would be highly correlated because they were expected to measure the same construct in different formats. In testing this hypothesis, a total of 14 participants (1 native English-speaking PhD student in applied linguistics, 6 Japanese PhD and MA students in applied linguistics, and 7 Japanese undergraduates studying law, literature, medicine, or technology) individually took the productive version with 30 randomly

selected items and then the receptive version with the same 30 items. For the productive version, they were asked to write answers in English or Japanese for the part of speech and the meaning sections and to underline a word or phrase for the contextual clue section. The responses in the productive version were scored by a native English teacher with a high proficiency in Japanese and the researcher (a native Japanese speaker). The part of speech items did not cause a problem with inter-rater reliability because it was easy to determine whether the answer was correct or wrong. For the contextual clue section, responses were regarded as correct if the word or phrase of the correct answer was included in the word or phrase the participants underlined. For the items in these two sections, correct answers were scored as correct (1) and wrong answers were scored as wrong (0). For the meaning section, responses were classified into correct, partially correct, and incorrect responses. Correct responses corresponded to the meaning and the part of speech of the original word. Partially correct responses had an incorrect but not totally wrong meaning (e.g., *sea* for the original word *seabed*). Correct responses were scored as correct (1), partially correct responses were scored as half correct (0.5), and wrong responses were scored as wrong (0). Inter-rater reliability as measured by Spearman's rank correlation coefficients[15] between the scores from the two raters was high (1.00 for the part of speech section, .97 for the contextual clue section, and .96 for the meaning section). For the productive version, average raw scores from the two raters were used for analysis. The responses on the items in the receptive version were scored either as 1 (correct) or 0 (wrong). Table 25 presents Spearman's rank correlation coefficients between the scores from these two versions.

[15] Spearman's ρ instead of Pearson's r was used for investigating the inter-rater reliability because the raw scores are ordinal.

Table 25. Correlation coefficients between the scores from the productive and the receptive versions

Section	Spearman's ρ
Part of speech	.91*
Contextual clue	.77*
Meaning	.81*

N=14; *p<.05.

The part of speech and the meaning sections had the correlation coefficients greater than .80, indicating a strong tendency that a person with a higher score on the productive version also got a higher score on the receptive version. A relatively low correlation coefficient (.77) was found in the contextual clue section, because one participant performed contrary to the other participants; that is, she got a higher score on the productive version than on the receptive version because she left some of the items unanswered in the receptive version. If this person was excluded from the analysis, Spearman's ρ increased to .89. This may serve as convergent evidence for the external aspect of construct validity.

Second, the correlations between GCT scores and self-reported TOEIC scores were examined. It was hypothesised that TOEIC and GCT scores would be positively correlated, but the correlations between TOEIC and GCT scores would be lower than the correlations between the scores from any two sections of the GCT (e.g., the scores from the meaning section of the GCT would be more closely related to the scores from the part of speech and the contextual clue sections of the GCT than those from TOEIC). TOEIC is a test of English reading and listening skills for business which may involve the skill of guessing from context as an important component. Some questions directly measure the meaning of difficult words. In many cases where test-takers come across unknown words, they may need to guess the meanings of the words in order to understand the reading and listening passages. For this reason, the correlations between

GCT and TOEIC scores (GCT-TOEIC correlations) were expected to be positive; however, these correlations were considered to be lower than the correlations between the scores from any two sections of the GCT (within-GCT correlations), because the GCT measures different aspects of the skill of guessing. This hypothesis was tested by examining whether the within-GCT correlations were higher than the GCT-TOEIC correlations. As this analysis was conducted based on 134 participants (31.3% of the participants who were administered the GCT) who reported their TOEIC scores, it was necessary to investigate the generalizability of the results from the 134 participants to the overall 428 participants. In so doing, the difference in Rasch person ability estimates between the 134 TOEIC score reporters and the others (294 non-reporters) was examined through Welch's *t*-test (2-tailed). As shown in Table 26, no statistically significant difference (α = .05) was found between the two groups for the three sections. The effect size (*r*) indices show small differences ($r < .2$) between the person ability estimates from the two groups for the three sections. This indicates that the 134 reporters may be representative of the overall 428 participants.

Table 26. Rasch person measures, *t*-statistics, and effect size between the reporters and non-reporters for the three sections

Section	Reporters		Non-reporters		Welch's	*d.f.*	*p*	*r*
	M	*S.D.*	*M*	*S.D.*	*t*			
Part of speech	1.80	1.79	1.66	1.58	0.82	231.4	.413	.054
Contextual clue	0.38	1.14	0.23	0.94	1.33	219.3	.186	.089
Meaning	0.06	1.09	-0.12	0.81	1.79	202.8	.076	.125

Given the representativeness of the 134 participants, a matrix of the Pearson's product-moment correlation coefficients between the GCT and TOEIC scores from the 134 participants is presented in Table 27.

Table 27. Correlation coefficients between GCT and TOEIC scores

	Part of speech	Contextual clue	Meaning
Contextual clue	.550*		
Meaning	.608*	.658*	
TOEIC	.239*	.295*	.463*

N=134; **p*<.05.

Table 27 shows that the GCT and TOEIC scores correlated positively (r = .239, .295, .463), but the GCT-TOEIC correlations were lower than the within-GCT correlations (r = .550, .608, .658). In order to determine whether there are statistically significant differences between these two groups of correlation coefficients (GCT-TOEIC vs. within-GCT correlations), a *Z*-test was performed for each of the three sections by means of a Meng-Rosenthal-Rubin method specifically designed for examining the difference between the overlapping correlation coefficients (correlation coefficients having a variable in common) (Meng, Rosenthal, & Rubin, 1992). Table 28 shows that for all three sections, the within-GCT correlations were significantly higher than the GCT-TOEIC correlations (α = .05). This shows that the above-mentioned hypothesis (positive correlations for the GCT-TOEIC scores but lower than within-GCT correlations) may be acceptable

Table 28. Difference between the within-GCT and the GCT-TOEIC correlations

Section	within-GCT correlations	GCT-TOEIC correlations	*Z*	*p*
Part of speech	r_{PC}=.550	r_{PT}=.239	3.40	.001
	r_{PM}=.608	r_{PT}=.239	4.70	.000
Contextual clue	r_{CP}=.550	r_{CT}=.295	2.75	.006
	r_{CM}=.658	r_{CT}=.295	4.84	.000
Meaning	r_{MP}=.608	r_{MT}=.463	2.18	.029
	r_{MC}=.658	r_{MT}=.463	2.51	.012

Note. *N*=134, P = part of speech section, C = contextual clue section, M = meaning section, T = TOEIC (e.g., r_{PC} = correlation coefficient between the scores of the part of speech section and the contextual clue section).

4.6.6 Consequential Aspect

The consequential aspect of construct validity "appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use" (Messick, 1995, p. 745). Providing evidence for this aspect is not an easy task.[16] Positive consequences such as improved language proficiency may be associated with support for test validity, but the improvement is usually affected by a number of other factors than the test quality such as the type of class activities, teaching methods, learning materials, and curriculum. That is, it is difficult to attribute positive consequences to the quality of the test alone. In terms of test validation, the primary concern is to minimise sources of invalidity such as construct-irrelevant and construct under-representation difficulty in order to avoid negative impact on the consequences resulting from the score interpretation and use (Messick, 1996). That is, low scores should not occur because the test is measuring something different from what it purports to measure or because the test fails to include important construct-relevant items that, if present, would allow the test-taker to achieve higher scores. As discussed in the content aspect of construct validity, the content relevance and representativeness of the GCT was supported by both theoretical argument and empirical evidence. This may be taken as supportive evidence for the consequential aspect.

Negative impact on consequences may also be caused by unfairness in test use (Messick, 1989, 1995, 1996). An unfair test gives a group of people an advantage over another. One way of evaluating this through Rasch measurement is to investigate item bias (Smith Jr., 2004b). Item bias refers to different item difficulties across groups of test-takers (Smith, 1992). As discussed in the generalizability aspect (item calibration

[16] The investigation of the use of the test by teachers in particular educational contexts would also be useful in evaluating the consequential aspect.

invariance), DIF analyses showed that the items did not work in favour of one group of test-takers against another in terms of gender and L1[17], indicating that the GCT is unlikely to cause negative consequences because of unfairness (item bias). Unfairness may also occur in scoring when the responses are graded subjectively by judges. The GCT does not cause this type of unfairness because it is written in a multiple-choice format which is free from variations in judge severity. This may be taken as supportive evidence for the consequential aspect of construct validity.

The consequential aspect may include washback, or the degree of behavioural and attitudinal change in teachers and learners with the introduction of a test; however, washback is only circumstantial to the validity argument, because even a poor test could conceivably cause positive washback if, for example, learners worked hard to prepare for the test regardless of its quality (Messick, 1996). In terms of validation, it is important to minimise sources of invalidity such as construct-irrelevance and construct under-representation difficulty so that negative washback may not occur (Alderson & Wall, 1993). Such invalidity was not observed for the GCT as mentioned above. Taken together, the GCT is unlikely to cause a negative impact on the score interpretation and use, which may be taken as supportive evidence for the consequential aspect of construct validity.

4.6.7 Responsiveness Aspect

Responsiveness, or sensitivity, refers to the degree to which an instrument is capable of detecting changes in person measures following an intervention that is assumed to impact the target construct (Medical Outcomes Trust Scientific Advisory Committee,

[17] With limited data on learners with different L1s, it did not work in favour of one group but further examination of this would be useful.

1995). This may be examined by examining a ceiling effect using a Rasch person-item map. A ceiling effect decreases responsiveness because able persons cannot demonstrate their gains from an experimental intervention such as teaching. As shown in the Rasch person-item maps in Figures 12 and 13, no ceiling effects were observed for the contextual clue and the meaning sections. This may be taken as supportive evidence for responsiveness. The person-item map in Figure 11, on the other hand, showed that the part of speech section may suffer from a ceiling effect. In fact, 35 (8.2%) participants got all items correct. This indicates that the part of speech section may not be sensitive enough to detect able persons' gains from an experimental treatment.

Responsiveness may also be investigated through Rasch person strata. As with item strata, a person strata index indicates the number of statistically different levels of person ability, and is derived by the following formula:

$$\text{Person strata} = (4\,G_{person}+1)/3,$$

where G*person* is Rasch person separation. With regard to the acceptable number of item strata, Wolfe and Smith Jr. (2007, p. 223) argue that "if the intended use of the measures is to distinguish between experimental and control groups, the assessment must be able to distinguish between at least two levels of trait (a person strata of 2)." Person strata are presented for each test form in the three sections in Table 29.

Table 29. Person strata for the three sections

	Form						Ave.
	A	B	C	D	E	F	
Part of speech	2.56	2.29	1.83	2.83	2.19	1.97	2.28
Contextual clue	1.95	2.29	2.00	2.44	2.19	2.29	2.19
Meaning	2.25	1.99	2.09	2.21	2.72	2.19	2.24

While the average person strata exceeded 2 for all three sections, four forms had the person strata of slightly smaller than 2 (1.83, 1.95, 1.97, and 1.99). This indicates that

these four forms may be problematic in terms of responsiveness. As low person strata indices are often caused by a small number of items (Linacre, 2010a), a successful attempt was made to solve this problem by creating new forms that include an increasing number of items so that each form would have person strata greater than 2 (See Section 4.7 for further discussion).

4.6.8 Interpretability Aspect

The interpretability aspect of construct validity refers to the degree to which qualitative meaning can be assigned to quantitative measures (Medical Outcomes Trust Scientific Advisory Committee, 1995). Rasch measurement provides a useful tool for interpreting the scores: a person-item map. The Rasch person-item map, which was presented in Section 4.6.2, expresses person ability and item difficulty estimates on a common interval scale, and shows the probability of a person's success on an item regardless of whether or not the person actually answered the item (missing data). At any point on the scale where a person and an item share the same location, the person has a 50% probability of getting the item correct. The lower an item is located, the higher success probability that person has on the item, and vice versa. This map facilitates the interpretation of the scores for the two main forms of assessment: norm-referenced and criterion-referenced. For norm-referenced interpretations, a person's score is compared to the scores of a group to see where the person's performance lies. The person-item map directly displays the location of a person with respect to the latent trait being measured; that is, more able persons are located towards the top and less able persons are located towards the bottom. Unlike the analysis based on raw scores which are ordinal, the person ability on the person-item map is displayed on an interval scale. This

means that one unit on the scale represents the same magnitude of the latent trait being measured across the whole range of the scale. For example, the difference between person abilities of 0 and 0.5 logits is the same as the difference between person abilities of 2.5 and 3.0 logits. For raw scores, on the other hand, the difference does not have any meaning. For example, the difference between persons who answered 90% correctly and who answered 100% correctly may not be the same as the difference between persons who answered 40% correctly and who answered 50% correctly. This indicates that norm-referenced interpretations are facilitated by the Rasch person-item map.

For criterion-referenced interpretations, cut scores are typically used to classify the test-takers into groups according to the level of performance, and then labels are given to these groups to provide meaning to the levels (e.g., advanced/intermediate/beginner). The person-item map directly indicates whether a person passes on a particular cut score. For example, if the cut score were set at an item difficulty of 1.5 logits, persons with an ability estimate of 1.5 logits or above could be regarded as passing on the cut score. This provides a clearly interpretable description of what a person can be expected to accomplish: a person with an ability estimate of 1.5 logits or above has a more than 50% probability of succeeding on an item with a difficulty estimate of 1.5 logits. Taken together, the use of a Rasch person-item map facilitates the interpretation of scores for both norm- and criterion-referenced assessments.

A difficulty in interpreting the results using a Rasch person-item map may derive from the unit of measurement (logit). Logit is a contraction of log-odds unit (of success). Odds are defined as the probability of success divided by the probability of failure. The natural logarithm of this odds ratio is called logit. For example, if the odds ratio of success and failure is 2:1, then the natural log of 2 equals 0.69 logits. If the odds ratio of

success and failure is 1:2, then the natural log of 0.5 equals -0.69 logits. As with all interval scales such as temperature, the origin of the scale is indeterminate; and thus the origin is usually set to the average item difficulty for convenience. The scale ranges from negative infinity to positive infinity.

Despite the usefulness of the Rasch person-item map, specialised computer software such as WINSTEPS is currently needed for obtaining the map. The simplest form of score reporting may be to use raw scores, because teachers and learners have only to count the number of correct responses and do not need to use specialised computer software. In order to investigate the adequacy of using raw scores for interpretation, Spearman's rank correlation coefficients between the raw scores and the Rasch person ability estimates were examined.[18] As shown in Table 30, the raw scores were highly correlated to the Rasch person ability estimates ($r > .9$) regardless of the use of a missing data design. This indicates that the total number of correct responses may serve as a close approximation to the latent trait of guessing from context. It should be noted here that the raw scores are only ordinal and are not on an interval scale. Thus, the difference between the score of 5 and 10 is not identical to the difference between the scores of 15 and 20.

Table 30. Correlation coefficients between the raw score and the Rasch person ability estimate for the three sections

Section	ρ
Part of Speech	.983*
Contextual clue	.990*
Meaning	.966*

$p<.05$

Table 31 presents the relationships between the raw scores and the Rasch ability

[18] Spearman's ρ was used because although the Rasch ability estimates were on an interval scale, raw scores were only ordinal and not suitable for Pearson's r.

estimates for the three sections. The raw scores were converted to the percentage of correctly answered items that the participants had actually taken.[19] This table shows that, for example, a person who got 80% of the items correct for the part of speech section has a Rasch ability of approximately 1.76 logits, which indicates that this person has a 50% probability of succeeding on a part of speech item with a Rasch difficulty estimate of 1.76 logits. This person has a greater probability of succeeding on any item with a Rasch difficulty estimate of less than 1.76 logits, and vice versa. If a cut score is set at a Rasch item difficulty of 1.5 logits, this person is taken as passing on the cut score.

Table 31. Conversion table of raw scores and Rasch ability estimates

Raw scores (%)	Part of speech	Contextual clue	Meaning
100	4.89	4.26	4.55
90	2.80	2.27	2.29
80	1.76	1.38	1.49
70	0.78	0.90	0.90
60	-0.02	0.43	0.36
50	-0.89	0.01	-0.01
40	-1.26	-0.43	-0.44
30	-1.85	-0.94	-0.90
20	-2.47	-1.50	-1.55

Taken together, the use of a Rasch person-item map guarantees a high degree of interpretability. Raw scores may also be used as a rough approximation of Rasch person ability estimates for convenience, but the interpretation needs to be made cautiously because the raw scores are only ordinal and the magnitude of difference between any two scores has no meaning.

This section has investigated the validity of the GCT from eight aspects of construct validity (content, substantive, structural, generalizability, external, consequential, responsiveness, and interpretability). The evidence provided in this

[19] The Rasch ability estimate for each raw score category was based on the persons whose scores ranged between ±1% of the category. For example, the Rasch ability estimate for the raw score of 90% was calculated by averaging the person abilities of those who got 89-91% of the items correct.

section generally indicates that the GCT is a highly valid measure for assessing the skill of guessing from context.

4.7 Creating New Forms

New test forms need be created for the GCT, because the six forms used for the present validation are different in the degree of reliability and in the number of items after the deletion of misfit items. This section explains how new forms were created and how the scores from the forms may be interpreted and reported to learners.

4.7.1 Equivalent Forms

At least two equivalent forms of the GCT are needed to serve as a tool for future research in this field. Equivalent forms have the same construct to be measured, the same test length, and the same distribution of item difficulties. Having two equivalent forms will allow a pre- and post-test design where the effects of teaching on the skill of guessing from context may be investigated.

The first step for creating new forms was to determine the number of items included in each form in order to achieve a certain level of reliability. The minimum level of reliability was determined so that the Rasch person strata indices would exceed 2. A Rasch person strata index of 2 indicates two statistically distinct levels for person abilities, which is the minimum level for acceptable responsiveness (detecting change after an experimental treatment). The person strata index of 2 is equivalent to person reliability of .610 given the formulae in Linacre (2010a).[20] The number of items needed for achieving the reliability of .610 was estimated based on the following Spearman-

[20] Reliability = $G^2/(1+G^2)$, and Strata = $(4G+1)/3$, where G = separation coefficient.

Brown prediction formula (Brown, 1910; Spearman, 1910):

$$T = C \times \frac{R_T\,(1 - R_C)}{R_C(1 - R_T)},$$

where T = target number of items, C = current number of items, R_T = target person reliability, and R_C = current person reliability. Table 32 shows the estimated number of items that are required to arrive at the person reliability of .610 for each form of the three sections.

Table 32. Estimated number of items needed for arriving at person strata of 2

	Form A	Form B	Form C	Form D	Form E	Form F
Part of speech	9.3	14.0	16.0	6.6	14.5	16.7
Contextual clue	18.5	19.0	13.0	9.6	16.5	11.8
Meaning	13.1	19.8	11.9	11.6	8.9	12.9

Form B in the meaning section indicates the largest number of items (19.8) for arriving at the Rasch person reliability of .610 (= Rasch person strata of 2). This means that a new test form should involve at least 20 items in order for any form to guarantee the minimum requirement for a sensitive test (Rasch person strata of 2). As indicated by the pilot studies (see Section 3.7), a 20-item test form may be completed in half an hour and is unlikely to result in a fatigue effect that could affect reliability. Thus, new test forms had 20 items which was the minimum number in terms of reliability and the maximum number in terms of fatigue effect.

As there are 49 acceptable items, two equivalent 20-item test forms can be constructed. The two equivalent forms were created based on the following criteria in order to maintain the representativeness of the construct:

1. Each form had nine nouns, six verbs, three adjectives, and two adverbs in order to reflect actual language use (noun: verb: adjective: adverb =

9:6:3:2).

2. Each form included all twelve types of contextual clues (one or two items per contextual clue) in order to ensure test representativeness.

3. The proximity of the clue to the test word was controlled so that each form had the same number of clue-inside (clues that appear in the same sentence as the test word) and clue-outside (clues that appear in a different sentence from the sentence containing the test word) items; that is, 13 clue-inside items and 7 clue-outside items for each form. This ratio (13:7) was an approximate ratio of 41:19 for the 60 original items (see Section 3.7).

4. In order to make sure that each form has items with a wide spread of difficulty, the 49 acceptable items were classified into four groups based on the item difficulties in the meaning section[21]: 1) larger than 0.5 logits, 2) between 0 and 0.5 logits, 3) between -0.5 and 0 logits, and 4) smaller than -0.5 logits. Each form had five items selected from each of the four groups.[22]

The distributions of the item difficulties for the two new forms for each section are shown in Figures 22-24 using the Rasch person-item map (The person ability and the item difficulty estimates larger than 2.0 and smaller than -2.0 are summarised into one row for want of space). The items of Form A are presented on the left-side of the item distribution, and the items of Form B are presented on the right-side of the item distribution. For each item, the item number is followed by its Rasch item difficulty in brackets. For example, 13(3.19) means that the item number is 13 and its item difficulty is 3.19 logits.

[21] The spread of item difficulties was determined based on the meaning section instead of the part of speech and the contextual clue sections, because deriving the meaning is arguably the most important aspect in the skill of guessing from context. As will be discussed later, however, no significant difference was found in item difficulty between the two forms for the part of speech and the contextual clue sections.

[22] To be precise, Form A had four items with difficulty estimates larger than 0.5 logits and six items with difficulty estimates between 0 and 0.5, because there were only a total of nine items with difficulty estimates larger than 0.5 logits.

<More able persons>			<More difficult items>							
			Form A				Form B			
	##########									
	##########									
	##########									
	######	S	13(3.19)				20(2.20)			
2	########	+					57(1.93)	5(1.89)		
	##########		32(1.53)				15(1.69)	33(1.66)		
	#########	M	S							
1	#########	+	59(0.93)	24(0.89)			1(0.93)			
	*##########		45(0.77)							
	######						36(0.61)	18(0.40)		
	####	S	17(0.28)	48(0.28)			28(0.28)			
0	*######	+	M	40(-0.04)						
	###		27(-0.13)	26(-0.27)	56(-0.33)		46(-0.22)	50(-0.22)	39(-0.24)	
	##		35(-0.37)	2(-0.38)			30(-0.27)			
	##		6(-0.71)	42(-0.85)			25(-0.73)	8(-0.84)		
-1	#		38(-0.91)	23(-1.02)	14(-1.07)					
	##	T	S	55(-1.35)				44(-1.15)	7(-1.17)	
	*##		10(-1.59)							
							12(-1.82)			
-2	*#	+								
	##*						34(-2.35)	53(-2.36)		
<Less able persons>			<Less difficult items>							

Figure 22. Person-item map of the equivalent forms for the part of speech section

<More able persons>			<More difficult items>							
			Form A				Form B			
	*#####									
2	*###	T	+							
	*#						5(1.69)			
	*#									
	*######	S	T				30(1.12)			
1	########	+	24(1.08)				28(1.08)			
	*########		14(0.69)							
	##########		S				34(0.50)			
	##########	M	2(0.33)	45(0.20)			1(0.26)	39(0.26)	46(0.15)	
0	##########	+	M	35(0.05)	6(-0.03)	40(-0.01)	33(0.12)	7(0.10)	18(-0.04)	
				59(-0.08)	13(-0.11)		20(-0.07)	50(-0.10)	8(-0.12)	
	*##########		17(-0.16)	10(-0.19)	23(-0.21)		15(-0.18)			
	########	S	32(-0.23)	42(-0.23)	56(-0.25)		53(-0.45)	12(-0.47)	25(-0.50)	
	####		38(-0.28)	55(-0.42)						
-1	#####	+	27(-0.69)	26(-0.83)	48(-0.91)		36(-0.89)	44(-0.91)	57(-0.94)	
	*##		T							
	##	T								
-2		+								
	*									
<Less able persons>			<Less difficult items>							

Figure 23. Person-item map of the equivalent forms for the contextual clue section

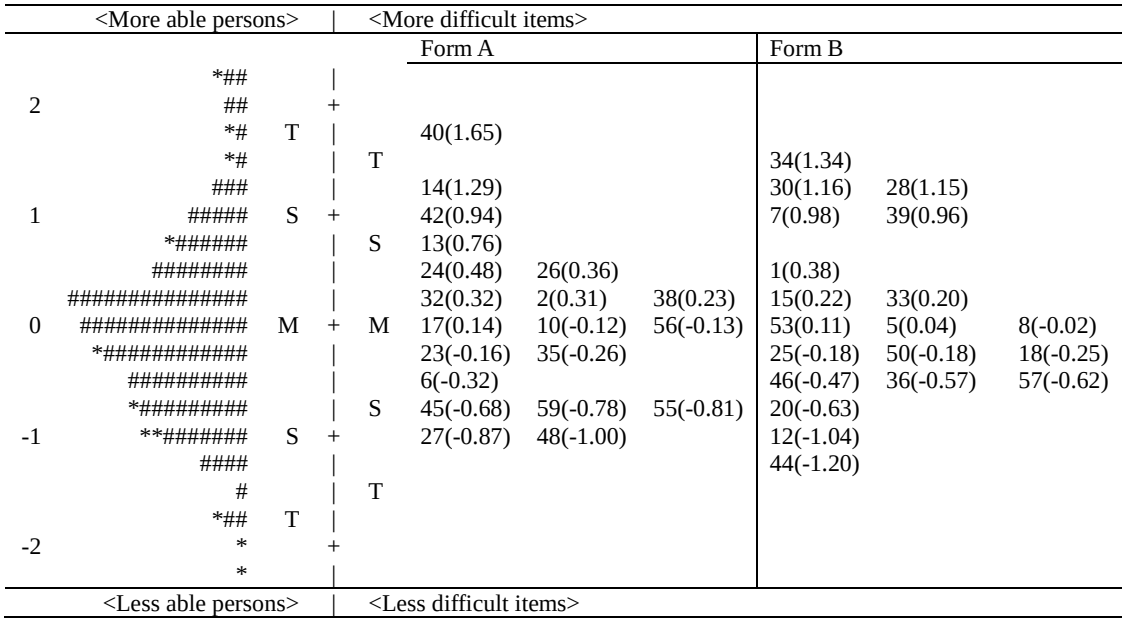


Figure 24. Person-item map of the equivalent forms for the meaning section

The person-item maps in Figures 22-24 indicate that the item difficulties are evenly distributed between Forms A and B for the three sections. In order to statistically examine the homogeneity of variance of item difficulty between the two forms, Levene's test was performed. The results showed that the null hypothesis of equal variances was not rejected for the three sections (F = 2.18, p = .148 for the part of speech section; F = 1.81, p = .187 for the contextual clue section; and F = 0.00, p = .957 for the meaning section), indicating that the spread of item difficulties may be acceptably equal between the two forms. Subsequent *t*-tests (2-tailed) did not detect any significant differences in the mean item difficulties between the two forms for any of the three sections (Table 33). The effect sizes (r) were smaller than .20, which indicates small differences between the two forms (Cohen, 1988, 1992).

Table 33. Comparison of the item difficulty between the two equivalent forms

	Form A		Form B		*t*	*d.f.*	*p*	*r*
	M	*SD*	*M*	*SD*				
Part of speech	-0.06	1.12	0.01	1.41	-0.17	38	.866	.027
Contextual clue	-0.11	0.46	0.03	0.69	-0.78	38	.440	.119
Meaning	0.07	0.73	0.07	0.74	-0.01	38	.995	.001

Taken together, the two forms may be representative of the construct being measured, and may be equivalent in the mean and the spread of item difficulties. (See Appendix F for the two new forms of the GCT.)

4.7.2 Score Interpretation

The interpretation of the scores from the new test forms may be facilitated by Rasch measurement output including Rasch person ability estimates and Rasch person-item maps. For norm-referenced interpretations, a Rasch person-item map may help visually understand a learner's performance within a particular group of people, because more able learners are located towards the top and less able learners are located towards the bottom. More precise information on learners' performance may be obtained from Rasch person ability estimates. As Rasch ability estimates are linear, the mean and the standard deviation have substantial meaning (raw scores are typically ordinal and it is difficult to interpret the meaning of the mean and the standard deviation). The use of Rasch ability estimates may also allow the comparison between multiple groups of learners and the investigation of the development of the skill of guessing from context through statistical tests such as a *t*-test and an ANOVA. When one is only interested in knowing the location of a learner's performance in a group, raw scores instead of Rasch person ability estimates may be used for convenience.

For criterion-referenced interpretations, cut scores may be predetermined so that

test-takers may be classified into groups according to their level of performance. The number of cut points may be determined based on item strata indices which indicate the number of statistically distinct levels. Given the item strata indices shown in Table 11 (6.07 for the part of speech section, 3.57 for the contextual clue section, and 4.85 for the meaning section), having three cut points (four levels) would be statistically justified for the part of speech and the meaning sections. For the contextual clue section, the item strata fell a little short of 4, but three cut points were also used, because the item strata approached 4 (3.57) and different cut points for different sections may make the score interpretation more complicated. The three cut points were set at 1, 0, and -1 logits to create four difficulty levels, because each section had items with difficulty estimates around 1, 0, and -1 logits (see Figures 22-24) and thus was considered to be useful in differentiating between the levels. The four levels are summarised in Table 34. For easier interpretation, the corresponding raw scores are also presented as a rough approximation. It should be noted here that as can be seen from Figure 23, the contextual clue section may be less sensitive to people at Level 1, because it does not have easy items (less than -1.0 logits).

Table 34. Levels for criterion-referenced interpretations

Level	Range	Label	Approximate raw score range		
			P	C	M
4	Above 1 logits	High	16-20	16-20	16-20
3	0 ~ 1 logits	Relatively high	13-15	11-15	11-15
2	-1 ~ 0 logits	Relatively low	10-12	6-10	6-10
1	Below -1 logits	Low	0-9	0-5	0-5

Note: P=part of speech section, C=contextual clue section, M=meaning section.

This section has provided a proposal for interpreting the scores obtained from the two new forms of the GCT. The subsequent section discusses how the scores may be presented to learners.

4.7.3 Score Reporting to Learners

For practical use of the GCT, diagnostic feedback needs to be easy for learners and teachers to understand so that learners' weaknesses in guessing from context may be clearly indicated. To meet this need, a bar graph may be useful because the information is visually presented and intuitively interpretable. For example, Learner A's estimated ability (P = 1.97 logits (Level 4), C = -0.23 logits (Level 2), M = -0.61 logits (Level 2))[23] is presented in a bar graph in Figure 25. The horizontal axis indicates the section of the GCT (PoS = part of speech section, Clue = contextual clue section, and Meaning = meaning section). The vertical axis indicates the level of the learner based on the criteria shown in Table 34. The bar graph shows that this learner demonstrated very good knowledge of part of speech (Level 4), but his performance on the contextual clue and the meaning sections was relatively low (Level 2); thus, his weakness lies in finding contextual clues (and deriving the meaning based on that information). The learner (or teacher) may then be able to prioritize the learning of contextual clues to potentially improve guessing (see Section 4.8 for further discussion).

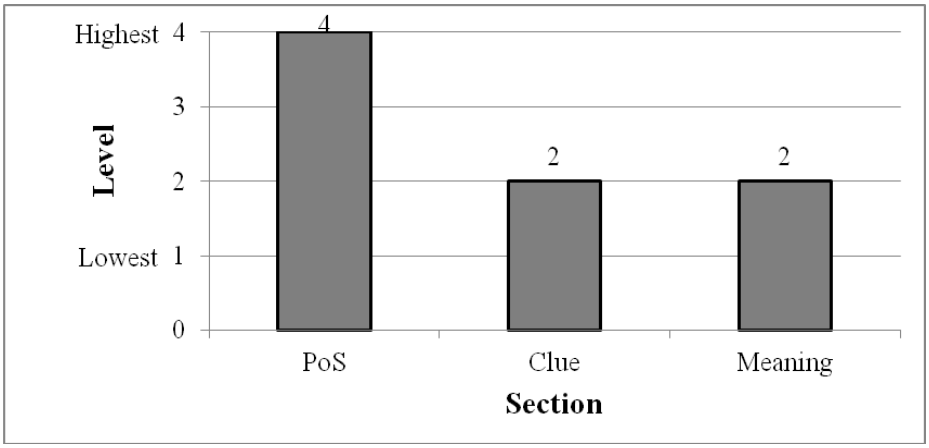


Figure 25. Score report (Learner A)

[23] The following abbreviations will be used in this section: P = Part of speech section, C = Contextual clue section, and M = Meaning section.

Another typical example may be seen in Learner B's estimated ability (P = -1.26 logits (Level 1), C = 0.35 logits (Level 3), M = -0.15 logits (Level 2)). This learner's performance is presented in Figure 26. This learner demonstrated relatively good knowledge of contextual clues, but her performance on the part of speech and the meaning sections was relatively low; thus, this learner's weakness lies in identifying the part of speech of unknown words (and deriving the meaning based on that information). This indicates that this learner's guessing skill may be improved with knowledge of part of speech.

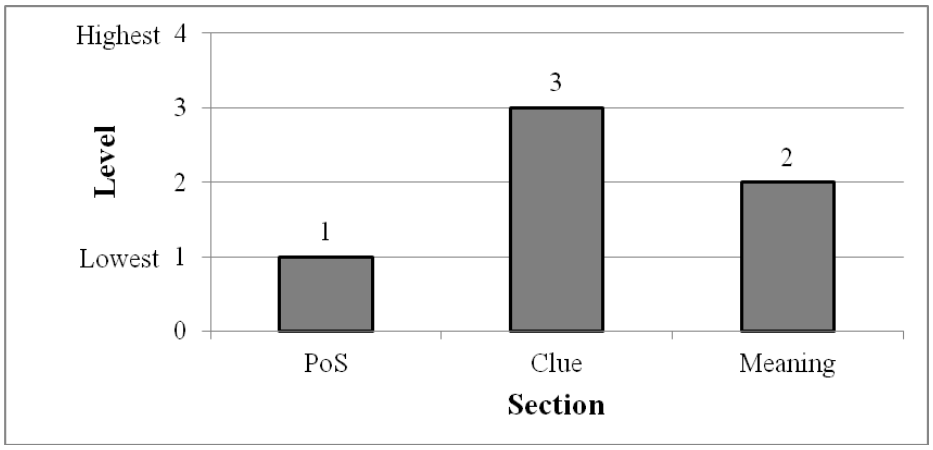


Figure 26. Score report (Learner B)

The bar graphs in Figures 25 and 26 were created based on the learners' Rasch person ability estimates (logit scores), but approximate results may be obtained conveniently based on raw scores with reference to the conversion table (Tables 31 and 34). For example, if a learner got the raw scores of P=17, C=16, and M=15, then this learner's levels are P=4, C=4, and M=3. If a learner got the raw score of 7 for each section, then this learner's levels are P=1, C=2, and M=2.

This section has looked at a practical application of the GCT; that is, creating two

new equivalent forms for investigating the development of the skill of guessing from context, interpreting the scores from the GCT, and reporting the scores to learners. The subsequent section discusses the theoretical values of the present research.

4.8 Discussion

Previous studies (Bruton & Samuda, 1981; Clarke & Nation, 1980; Williams, 1985) proposed guessing-from-context strategies from a pedagogical perspective, and these strategies essentially included the three aspects measured by the GCT. However, no attempts have been made to empirically examine whether identifying the part of speech of the unknown word and finding a contextual clue really contribute to deriving its meaning. As the present research measured the three aspects of guessing from context using the GCT, this issue may be addressed by investigating the interrelationships among the three aspects. In so doing, a multiple regression analysis was performed with the dependent variable being the Rasch person ability estimates from the meaning section and the independent variables being the Rasch person ability estimates from the part of speech section and from the contextual clue section. A path diagram of the results is presented in Figure 27 which is the same as the one presented in Figure 18. This figure shows that both the ability to identify the part of speech of the unknown word (β = .32) and the ability to find a contextual clue (β = .44) significantly contribute to the ability to derive its meaning. A combination of the abilities of part of speech and contextual clues accounted for about half of the variability of the ability to derive the meaning (R^2=.45). Given that guessing involves many other factors such as reading ability and world knowledge, this coefficient of determination may be considered high. Taken together, the results showed that both identifying the part of speech of the

unknown word and finding a contextual clue to help guess meaning play an important role in deriving meaning, indicating the effectiveness of the guessing strategies proposed by the previous studies.

One of the important features of the GCT is its comprehensiveness (measuring multiple aspects of the skill of guessing). This is in line with recent L2 vocabulary studies (Schmitt, 1998; Webb, 2005, 2007a, 2007b, 2007c) which have underscored the importance of measuring multiple aspects of vocabulary knowledge because different tasks may have varying effects on aspects of vocabulary knowledge. The GCT presupposes that different tasks and teaching materials may result in the development of different aspects of guessing skill and the ability of deriving meaning. For example, the instruction of contextual clues may improve scores on the contextual clue section and lead to the improvement of the ability of deriving meaning. Grammar instruction may improve scores on the part of speech section and contribute to the improvement of the ability of deriving meaning. The introduction of the guessing strategies may also raise learners' awareness of the importance of identifying the part of speech and looking for contextual clues, which may lead to an improvement in guessing. The GCT may thus contribute to effective and efficient teaching of the skill of guessing from context.

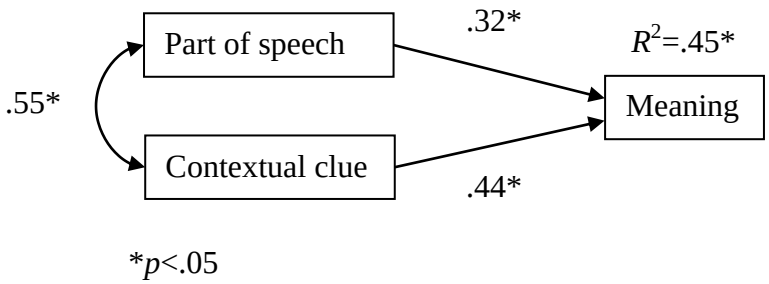


Figure 27. Relationships of the part of speech and the contextual clue sections to the meaning section

4.9 Summary

This chapter aimed to validate the GCT so that it would be widely available to researchers, teachers, and learners. In so doing, 428 Japanese learners of English with a wide range of proficiency levels participated in the present research. Six forms each with 20 items were created in a paper-based format and were randomly distributed to the participants. Rasch analysis showed that lucky guessing (unexpected success on difficult items by persons with low ability) was observed in the part of speech section, but not in the contextual clue and the meaning sections; and thus, the responses in the part of speech section were corrected for lucky guessing. Rasch analysis also revealed that 49 out of 60 items were acceptable. The validity of the GCT with the 49 acceptable items was investigated from eight aspects of construct validity (content, substantive, structural, generalizability, external, consequential, responsiveness, and interpretability) in order to comprehensively provide logical argumentation and empirical evidence to support its validity. Table 35 summarises the evidence provided for the validity argument. On the whole, both the logical argumentation and the empirical evidence indicated a high degree of validity. The validity argument also revealed the following three points to note:

1. Further evidence may still be needed for item calibration invariance and person measure invariance in the generalizability aspect because the small sample size may have affected the results.

2. The part of speech section may not be responsive (or sensitive) enough to detect able persons' gains from an experimental intervention because of a ceiling effect.

3. Five items with unacceptable Rasch measurement statistics need watching for future use of the GCT. A close look at the passages and the options did not find any problems with these items. Whether these items work well needs to be examined further. Table 36 presents the five items that need watching.

Two equivalent forms were created so that each form had 20 items with a wide spread of difficulty. Each form had 20 items so that any form would have person strata of greater than 2 which is the minimum requirement for a responsive test. These new forms are useful for research involving a pre- and post-test design. The new forms are also useful for teachers and learners because the results may provide learners with diagnostic feedback on their weaknesses in guessing from context. The scores obtained from the GCT are highly interpretable for both norm- and criterion-referenced purposes in the context of Rasch measurement. For more convenient interpretations, conversion tables (Tables 31 and 34) between raw scores and Rasch person ability estimates are provided. The scores may be effectively reported to learners using a bar graph which presents learners' weaknesses visually. Taken together, it should be reasonable to conclude that the GCT is a highly valid measure for assessing the skill of guessing from context and useful for both research and practical purposes.

Table 35. Summary of evidence provided for the GCT

Aspects	Sub-category	Evidence provided
Content	1. Content relevance	Test specifications
	2. Representativeness	Rasch item strata
		Rasch person-item map
	3. Technical quality	Rasch item fit analysis
	4. Expert judgments	Reviews by English teachers and PhD students in applied linguistics
Substantive		Test of difficulty hypotheses
		Rasch person fit analysis
Structural		Dimensionality analysis
Generalizability	1. Item calibration invariance	DIF analysis for gender and L1
	2. Person measure invariance	DPF analysis for item order
	3. Reliability	Rasch person separation and reliability
		Rasch item separation and reliability
	4. Invariance across administrative contexts	Comparison between person ability estimates from a 20-item form and the missing data design
External		Correlation with the productive version of the GCT
		Correlation with TOEIC scores
Consequential		Analysis of sources of invalidity
		Item bias
Responsiveness		Person-item map (ceiling effects)
		Person strata
Interpretability		Person-item map
		Conversion of raw scores and Rasch person ability estimates

Table 36. Summary of items that need inspecting for future use of the GCT

Item No.	Section	Reason for future inspection
5	M	Technical quality (underfit)
57	M	Technical quality (underfit)
24	P	DIF analysis (gender)
58	C	DIF analysis (gender)
30	M	DIF analysis (gender)

Note: M=meaning, P=part of speech, C=contextual clue

CHAPTER 5

DEVELOPMENT OF THE WORD PART TEST

This chapter describes the procedure for developing a word part test. Beginning with the purpose of the test, it presents the approach to selecting word parts for the test and examines the quality of the selected word parts. It also discusses which test format is most useful.

5.1 Purpose

The word part test (WPT) aims to measure L2 learners' comprehensive written receptive knowledge of English word parts. Receptive word part knowledge refers to being able to recognise the form, the meaning, and the function of word parts. Productive word part knowledge, on the other hand, refers to being able to spell and pronounce word parts correctly and use them to express their meaning or function. The WPT focuses on receptive knowledge instead of productive knowledge, because recognising word parts within words (receptive knowledge) may be related to VLP more closely than using correct word parts (productive knowledge). For example, recognising the word part *-ment* in *movement* may make it easier to learn the word *movement* than knowing that the verb *move* is nominalised with *-ment* and not *-ness*. In addition, receptive knowledge is easier to complete and grade. Productive knowledge has been measured typically by asking test-takers to write each part of speech for a word by adding derivational suffixes (Ishii & Schmitt, 2009; Schmitt, 1999; Schmitt & Zimmerman, 2002). For example, learners are asked to change the target word *assume* into a noun, a

verb, an adjective, and an adverb. This type of test takes more time to complete and grade.

5.2 Selection of Word Parts

In the present research, a word part refers to an affix, a bound morph which co-occurs with bases which contain free morphs (Bauer, 1983). A free morph can stand on its own, whereas a bound morph cannot. For example, *un-* which means 'not' is an affix, because it must attach to free morphs such as *happy* and *lucky* to form the words *unhappy* and *unlucky*. *Se-* which means 'apart from', on the other hand, is not an affix, because it attaches to bound morphs such as *clude* and *parate*, forming the words *seclude* and *separate*.

Affixes are roughly divided into two types: inflectional and derivational. Inflectional affixes indicate grammatical information such as case, number, and tense, rather than changing the meaning or the syntactic category of the bases. For example, the verb *walk* is marked by the inflectional affix *-s* to indicate the third person singular (*walks*), but *-s* does not change the meaning or the part of speech of *walk*. Derivational affixes, on the other hand, change the meaning or the part of speech of the word bases. For example, the addition of *-ness* to the adjective *kind* results in the noun *kindness*. The present research focuses on derivational affixes instead of inflectional affixes. By definition, an inflectional affix does not change the meaning or the syntactic property of the word base, resulting in words with the same meaning or the same syntactic property as the base. Inflectional affixes may not contribute to increasing knowledge of vocabulary to the same extent as derivational affixes, and thus do not meet the purpose of the WPT which regards affix knowledge as a facilitative factor for vocabulary

learning. The inflectional affixes involve plural, third person singular present tense, past tense, past participle, *-ing*, comparative, superlative, and possessive, which correspond to the affixes classified as Level 2 (inflectional suffixes) in Bauer and Nation's (1993) levels of affixes. [24]

Affixes were chosen from the most frequent 10,000 word families in the British National Corpus (BNC) word lists developed by Nation (2004) (available at http://www.victoria.ac.nz/lals/staff/paul-nation.aspx). In the BNC word lists where the word family is used as the unit of word counting, words are ordered in terms of their frequency in the BNC. The most frequent 10,000 word families could be taken as one of the goals for vocabulary learning, because knowledge of the most frequent 8,000-9,000 words is needed for comprehension of authentic written text (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006). It is likely that knowledge of these words makes it possible to understand the majority of written texts. Given that affix knowledge facilitates vocabulary learning, learning the affixes that appear in the most frequent 10,000 words should be a learning goal.

For the WPT, affixes that appear in more than one word family in the most frequent 10,000 word families of the BNC word lists were selected. For example, *un-* was chosen, because it appears in multiple word families such as *unhappy* and *unlucky* from these lists. *Ante-*, on the other hand, was not included, because it does not appear in these 10,000 word families. Words such as *anteroom* and *antenatal* are beyond the

[24] It should be noted here that morphologists do not agree as to which affixes are categorised as inflectional affixes. For example, Beard (1982) takes plural as derivational, whereas Carstairs-McCarthy (2002) takes it as inflectional. Moreover, the *-ing* forms and past participles may not be clearly inflectional. For example, *walking* is more likely to be inflectional in a sentence such as *He is walking now*, whereas it is more likely to be derivational in a sentence such as *Regular walking keeps you healthy*. Similarly, *excited* in the sentence *The performance of the rock group has excited the audience* is more likely to be inflectional, whereas in the sentence *The excited crowd burst into the street* it is more likely to be derivational.

10,000-word level. Affixes that appear only in one word family in the first 10,000 word families were not included, because they may not facilitate the learning of these 10,000 word families. For example, *di-* appears only in the word *dioxide* in the 10,000 word families. Even if learners find that *dioxide* can be broken into *oxide* and *di-* which means 'two', they have no chance of utilising the knowledge of *di-* to learn other words in the most frequent 10,000 word families.

Allomorphs (a morph which varies in spelling or sound but not in meaning) were treated as different word parts. For example, *im-* in *impossible* is an allomorph of *in-* in *informal*, both of which have the different spellings according to the subsequent sound but have the same meaning 'not'. These two affix forms were examined individually on the test, because knowledge of one does not necessarily guarantee knowledge of the other.

In summary, the selected word parts meet the following three criteria: (1) They are bound morphs which attach to free morphs; (2) They appear in more than one word family in the most frequent 10,000 BNC word families; and (3) Allomorphs are treated as different affixes. Using these criteria, a total of 118 affixes (42 prefixes and 76 suffixes) were identified (see Appendix B for a list of these affixes).

5.3 Quality of the Selected Affixes

This section aims to describe the quality of the 118 selected affixes by comparing them with the affixes selected in previous studies. More specifically, it investigates to what extent the 118 affixes cover the affixes included in other affix lists.

Several studies investigated the frequencies of affixes in The Teacher's Word Book of 20,000 Words (Thorndike, 1932) and The Teacher's Word Book of 30,000 Words

(Thorndike & Lorge, 1944) to create useful affix lists for learners. Thorndike (1941) investigated the usefulness of 90 suffixes by analysing the most frequent 20,000 words (Thorndike, 1932) in terms of frequency, ease of recognition, and ease of inferring meaning, and chose 24 suffixes that should be learned by children under grade 10. The WPT includes all of these suffixes. Stauffer (1942) identified 61 prefixes that appeared in the 20,000 words in Thorndike (1932) for the purpose of providing useful prefixes to teach in elementary schools. The affixes in the WPT cover only 41.0% of the prefixes in Stauffer's list. The reason for the low coverage is that Stauffer included infrequent prefixes such as *ante-* (*antechamber*) and prefixes that do not attach to free morphs such as *ad-* (*advert*). Bock (1948) identified 97 Latin affixes (42 prefixes and 55 suffixes) that appeared in the 20,000 words in Thorndike's word book. Coverage of the affixes in the WPT was 47.6% for the prefixes and 50.9% for the suffixes. The low coverage was due to the same reasons as Stauffer's list: the inclusion of infrequent affixes and affixes with bound morphs. Harwood and Wright (1956) identified the most frequent 32 suffixes in the 30,000 words in Thorndike and Lorge's (1944) word book. The WPT includes all of these suffixes.

Some studies examined affixes from various other perspectives. Bauer and Nation (1993) classified affixes into seven levels (Table 37) based on the eight criteria described in Table 38. The WPT covered 80 out of 81 affixes (24 prefixes and 57 suffixes) in Levels 3-6.[25] The only prefix excluded was *ante-* which did not appear in the most frequent 10,000 word families in the BNC word lists.

[25] Levels 1, 2, and 7 were excluded from the analysis, because Level 1 did not include any affixes, Level 2 involved inflectional suffixes which were not dealt with in the present research, and for Level 7 no comprehensive list of affixes was provided although some example affixes were discussed.

Table 37. The seven levels of affixes in Bauer and Nation (1993)

Level	No. of prefixes	No. of suffixes	Description
1	-	-	Each form is a different word
2	-	8	Inflectional suffixes
3	2	6	The most frequent and regular derivational affixes
4	1	10	Frequent, orthographically regular affixes
5	19	31	Regular but infrequent affixes
6	2	10	Frequent but irregular affixes
7	-	-	Classical roots and affixes

Table 38. The eight criteria for affix classification in Bauer and Nation (1993)

Criteria	Description
Frequency	The number of words in which an affix occurs.
Productivity	The likelihood that the affix will be used to form new words.
Predictability	The degree of predictability of the meaning of the affix.
Regularity of the written form of the base	The predictability of change in the written form of the base when the affix is added.
Regularity of the spoken form of the base	The amount of change in the spoken form of the base when the affix is added.
Regularity of the spelling of the affix	The predictability of written forms of the affix.
Regularity of the spoken form of the affix	The predictability of spoken forms of the affix.
Regularity of function	The degree to which the affix attaches to a base of known form-class and produces a word of known form-class.

More recently, Nation (2001) created a sequenced list of derivational affixes for learners of English, which categorised 97 affixes (35 prefixes and 62 suffixes) into five stages (Table 39). The WPT covers all the suffixes and 77.1% (27/35) of the prefixes in Nation's affix list. The eight excluded prefixes were infrequent (*ante-* (*antenatal*)) or did not attach to free morphs (*ad-* (*advert*), *com-* (*combine*), *ex-* (*exclude*), *in-* (*include*), *ob-*

(*obstruct*), *per-* (*percolate*), *pro-* (*proceed*)).

Table 39. Five stages in Nation's (2001) sequenced list of affixes

Stage	No. of prefixes	No. of suffixes	Description	Source
1	2	6	The most frequent and regular derivational affixes	Level 3 in Bauer and Nation (1993)
2	1	10	Frequent, orthographically regular affixes	Level 4 in Bauer and Nation (1993)
3	19	31	Regular but infrequent affixes	Level 5 in Bauer and Nation (1993)
4	2	10	Frequent but irregular affixes	Level 6 in Bauer and Nation (1993)
5	11	5	Classical roots and affixes	Stauffer (1942) Bock (1948) Harwood & Wright (1956)

A number of studies selected affixes for the purpose of measuring affix knowledge based on different criteria. In examining the acquisition of English derivational morphology by L1 children, Freyd and Baron (2000) selected 30 derived words by taking one word from each thousand words up to 30,000 words in the American Heritage Word Frequency Book (Carroll, et al., 1971). The 30 derived words had 20 different suffixes. The WPT covers 90% (18/20) of these suffixes. The two suffixes that were not included were *-ed* (*disordered*) and *-itude* (*servitude*). This was because *-ed* was an inflectional ending, and *-itude* did not attach to free morphs in the most frequent 10,000 word families in the BNC word lists.

In investigating the acquisition of morphology by L1 children, Tyler and Nagy (1989) classified derivational suffixes into two types: neutral and nonneutral. Neutral suffixes (e.g., *-er* and *-ness*) attach to independent words such as *teach* and *kind*, forming the words *teacher* and *kindness*. They also cause no change in pronunciation of

the word to which they are attached. For example, the addition of *-er* does not change the pronunciation of *teach* in *teacher*, and the same is true for *kindness.* Finally, neutral suffixes usually produce semantically transparent words. For example, *-er* means 'person' in words such as *teacher* and *sender*, and *-ness* means 'state/quality' in words such as *kindness* and *darkness*. Nonneutral suffixes (e.g., *-ify* and *-ity*), on the other hand, often attach to bound morphemes such as *qual*, forming *qualify* and *quality*. They tend to cause changes in pronunciation of the word to which they are attached. For example, *similar* is different in stress from *similarity*. Finally, nonneutral suffixes often produce words whose semantic relationships are not clear such as *native* and *nativity*. For their experiment, Tyler and Nagy selected a total of 24 derivational suffixes, eight suffixes each from nonneutral (*-ate, -ation, -atory, -ian, -ic, -ity, -ive,* and *-ous*), neutral low-frequency (*-age, -dom, -eer, -hood, -like, -ship, -some,* and *-wise*), and neutral high-frequency suffixes (*-ful, -ish, -ist, -ize, -less, -ly, -ness,* and *-s*). The WPT covers 22 out of 24 suffixes selected by Tyler and Nagy. The two suffixes that were not included were *-like* and *-s*. The reason for the exclusion of *-like* was that it was a free rather than a bound morph. The word *like* is used on its own as a preposition and is essentially similar in meaning to the suffix *-like*. For example, *childlike* means 'like a child' and *dream-like* means 'like a dream'. It seems that *-like* is similar in structure more to words that can be used in compounds such as *paper* in *newspaper* and *sandpaper*, than to suffixes such as *-ness* in *kindness* which do not stand on their own.[26] Although some morphologists (e.g., Beard 1982) argue that the plural *-s* is derivational rather than inflectional, the present research regards it as inflectional because it does not affect the meaning or the syntactic property of words. For example, *book* and *books* have the same

[26] Theoretically, *-like* is often categorised as a semi-suffix which stands between a suffix and a compounding element (e.g., Marchand, 1960). Other semi-suffixes include *-free* in *smoke-free* and *-worthy* in *trustworthy*.

meaning and the same syntactic property but differ only in number.

In an attempt to investigate the development of L2 learners' grammatical suffix knowledge, Schmitt and Meara (1997) dealt with the most common 14 suffixes that could attach to verbs (*-able*, *-age*, *-al*, *-ance/ence*, *-ed*, *-ee*, *-er/or*, *-ing*, *-ion*, *-ive*, *-ly*, *-ment*, *-s*, and *-ure*). Among them, the three inflectional suffixes *-ed*, *-ing*, and *-s* were not included in the WPT.

In examining Japanese learners' English affix knowledge, Mochizuki (1998) selected 26 prefixes and 56 suffixes that were supposed to be familiar to learners. The WPT covers 69.2% (18/26) of the prefixes and 98.2% (55/56) of the suffixes. The seven excluded prefixes included prefixes that did not attach to free morphs (*ambi-* (*ambidextrous*), *ana-* (*anachronism*), *com-* (*combine*), and *contra-* (*contradict*)), prefixes that occurred only once in the most frequent 10,000 word families in the BNC word lists (*extra-* (*extracurricular*)), and prefixes that were not bound morphs (*over-* (*overwork*) and *under-* (*understatement*)). [27] The only suffix excluded was the inflectional ending *-ed*.

In exploring the affix acquisition order for Japanese learners of English, Mochizuki and Aizawa (2000) selected 13 prefixes and 16 suffixes that met the following two criteria: (1) affixes in Levels 3-6 in Bauer and Nation's (1993) list of affixes, and (2) affixes used in more than two words in Nation's (1996) vocabulary list. The WPT covered 92.3% (12/13) of the prefixes and all of the suffixes. The only affix excluded was *ex-* in words such as *export* and *exclude* which did not attach to free morphs in the most frequent 10,000 word families in the BNC word lists.

Most of the affixes used in other studies with L1 children and L2 learners (e.g.,

[27] Theoretically, *over-* and *under-* are often categorized as compounding elements rather than prefixes (e.g., Bauer, 1983; Marchand, 1960).

Carroll, 1940; Nagy, Diakidoy, & Anderson, 1993; Wysocki & Jenkins, 1987) were covered by the WPT. The affixes that were not included were inflectional suffixes (e.g., *-ing*), infrequent affixes (e.g., *ante-*), and affixes attaching to bound morphs (e.g., *ad-*). (See Appendix C for a full list of affixes that were included in previous studies but not in the WPT).

In summary, affixes that were not included in the present affix list are categorised into four types: (1) inflectional suffixes (e.g., *-ed*, and *-s*); (2) infrequent affixes that are outside the most frequent 10,000 word families in the BNC word lists (e.g., *ante-* and *extra-*); (3) affixes that do not attach to free morphs (e.g., *ad-* and *com-*); and (4) affixes that can stand on their own (e.g., *over-* and *-like*). Coverage of the affixes in the previous studies by the WPT is summarised in Table 40. Coverage of prefixes was not high, mainly because many previous studies included prefixes that did not attach to free morphs (e.g., *ad-* and *com-*). Coverage of suffixes was lowered mainly by the exclusion of inflectional suffixes, but was high enough to approach 100%.

5.4 Aspects of Affix Knowledge

A number of researchers have proposed categories for what is involved in knowing affix knowledge. In an attempt to grasp the overall picture of L1 children's ability to understand and interpret written novel suffixed words, Tyler and Nagy (1989) argued that full knowledge of derivational suffixes involved the following three aspects: relational, syntactic, and distributional knowledge. Relational knowledge is the ability to recognise that two words are morphologically related, sharing a common meaning. For example, this knowledge makes it possible to see the morphological relationship between *create* and *creator* but not between *me* and *meter*. Syntactic knowledge involves knowing the syntactic property of a derivational suffix. For example, this

Table 40. Summary of coverage by the WPT

	Prefix			Suffix		
	No. selected*	No. covered**	% covered	No. selected*	No. covered**	% covered
Bauer & Nation (1993)	24	23	95.8%	57	57	100%
Bock (1948)	42	20	47.6%	55	28	50.9%
Carroll (1940)	20	9	45.0%	10	8	80.0%
Freyd & Baron (1982)	-	-	-	20	18	90.0%
Harwood & Wright (1956)	-	-	-	32	32	100%
Mochizuki (1998)	26	19	73.1%	56	55	98.2%
Mochizuki & Aizawa (2000)	13	12	92.3%	16	16	100%
Nagy, et al. (1993)	-	-	-	10	10	100%
Nation (2001)	35	27	77.1%	62	62	100%
Schmitt & Meara (1997)	-	-	-	14	11	78.6%
Stauffer (1942)	61	25	41.0%	-	-	-
Tyler & Nagy (1989)	-	-	-	24	22	91.7%
Wysocki & Jenkins (1987)	-	-	-	12	11	91.7%

Note * Number of affixes selected for each study; ** Number of affixes covered by the present affix list.

knowledge refers to knowing that *-ize* in *regularize* has the function of making a verb. Finally, distributional knowledge involves knowing the constraints on which suffixes can attach to bases with a certain syntactic property. For example, this knowledge

involves knowing that *-ness* attaches to adjectives (*quickness* is appropriate because *quick* is an adjective) but not to verbs (*playness* is not appropriate because *play* is a verb).

For the purpose of providing suggestions for determining the level of affixation of words in a word family, Bauer and Nation (1993) argued that four different types of knowledge were involved in being able to recognise the relationships between words in a word family. The first type is knowledge of word bases. For example, in order to be aware of the relationship between *kind* and *kindness*, learners need to know the word base *kind*. The second type involves being able to recognise known bases in words. For example, learners need to recognise that *kind* and *kindness* are related to each other because they share a common base *kind*. Thirdly, learners need to know the meanings or the syntactic properties carried by the affixes. For example, it is important to know *-less* means 'without' and makes an adjective. The last type involves being able to produce allowable base-affix combinations. For example, *-ness* attaches to adjectives, so that *kindness* is appropriate because *kind* is an adjective, but *moveness* is not appropriate because *move* is not an adjective.

In a discussion of how to use word parts for learning, Nation (2001, pp.275-278) argues that there are four aspects of knowledge required to use word parts. The first aspect is being able to recognise word parts in words. For example, learners need to be able to recognise that *unhappiness* consists of *un*, *happi*, and *ness*, each of which occurs in words such as *unpleasant*, *happily*, and *sadness*. The second aspect is being able to recognise the meaning or the syntactic property carried by an affix; for example, knowing that *-less* means 'without' and has the function of making an adjective. The third aspect is being aware of the changes of written and spoken form that occur when

an affix is added to a word. For example, learners need to be aware that when the suffix *-ion* is attached to *permit*, *t* in *permit* changes into *ss* and *permission* results. The last aspect involves knowing which classes of stems can take certain affixes. For example, *-ness* attaches to adjectives, so that *kindness* is appropriate because *kind* is an adjective, but *moveness* is not appropriate because *move* is not an adjective.

Regardless of different purposes for defining affix knowledge and different labels for categories, it seems that there is overlap between the studies. Relational knowledge in Tyler and Nagy's (1989) terminology is equivalent to the second type of knowledge discussed in Bauer and Nation (1993) (recognition of word bases). Relational knowledge is also analogous to the first aspect of word part knowledge mentioned in Nation (2001) (recognition of word parts), although the focus is different between the studies (word bases for Tyler and Nagy and Bauer and Nation, and word parts for Nation). Syntactic knowledge in Tyler and Nagy's terminology is included in Bauer and Nation's third and Nation's second aspects (knowledge of meaning and syntactic property). Finally, the three studies propose that knowledge of constraints on use of affixes plays a part in affix knowledge (Tyler and Nagy's distributional knowledge, Bauer and Nation's fourth aspect, and Nation's fourth aspect). The overlap of the aspects of affix knowledge among the three studies is summarised in Table 41. The overlapping categories are placed in the same line (e.g., "Relational knowledge" in Tyler and Nagy, "Recognition of word bases" in Bauer and Nation, and "Recognition of word parts" in Nation all share the same concept, and are thus categorised as Type 2.). The present research focuses on Types 2 and 3, which relate to receptive knowledge of affixes (recognition of affixes and their meanings and functions). Type 1 is knowledge of bases, rather than affixes. Types 4 and 5 relate to productive use of affixes which are

not dealt with in the present research.

Table 41. Types of affix knowledge

	Tyler & Nagy (1989)	Bauer & Nation (1993)	Nation (2001)
Type 1	-	Knowledge of word bases	-
Type 2	Relational knowledge	Recognition of word bases	Recognition of word parts
Type 3	Syntactic knowledge	Knowledge of meaning and syntactic property	Knowledge of meaning and syntactic property
Type 4	-	-	Knowledge of changes in spelling and pronunciation
Type 5	Distributional knowledge	Ability to produce allowable base-affix combinations	Knowing allowable affixes for stems

For receptive knowledge of affixes (Types 2 and 3 in Table 41), the three studies essentially included three types of knowledge: recognition of affix forms, their meanings, and their syntactic properties. It should be noted here that Tyler and Nagy did not include knowledge of affix meanings in their definition of knowledge of derivational morphology because their focus was on English derivational suffixes which typically change the syntactic categories of the bases.

5.5 Test Format

This section outlines the organisation of the WPT and the general format common to all items. It also determines the test format appropriate for measuring each aspect of the receptive affix knowledge (recognition of affix forms, their meanings, and their syntactic properties).

5.5.1 General Format

The WPT measures the three aspects of receptive affix knowledge individually: form, meaning, and use. The form section measures the ability to recognise written affix forms. The meaning section measures knowledge of affix meanings. The use section measures knowledge of syntactic properties carried by affixes. The terminology of the three sections (form, meaning, and use) is analogous to Nation's (2001) definition of what is involved in knowing a word. He argues that knowledge of words can be classified into form, meaning, and use at the most general level, each being subdivided into three (form involves spoken and written forms and word parts, meaning involves form-meaning relationships, concepts and referents, and associations, and use involves grammatical functions, collocations, and constraints on use). The form section focuses on written forms of affixes, the meaning section focuses on the relationships between affix forms and their meanings, and the use section focuses on grammatical functions of affixes.

To get an accurate assessment of receptive affix knowledge, all three aspects need to be measured. It is not reasonable to assume that learners acquire all three aspects of affix knowledge at the same time. A learner may be able to recognise *-y* in *difficulty* and see the semantic relationship between *difficult* and *difficulty*, but not know the function that *-y* makes. In a study with L1 children, Tyler and Nagy (1989) found that children appeared to develop relational knowledge (ability to recognise the relationship between a word and its derivative) before fourth grade, while knowledge of syntactic properties appeared to increase through eighth grade.

The test items were written using a multiple-choice format and no *Don't know* options were provided for the same reasons as the GCT (see Section 3.6.1). The WPT

involved four- instead of three-option items. Rodriguez (2005) argues that three options are optimal for the multiple-choice format, but the main reason for the preference of three options is that more three-option items can be administered per unit of time than four-option items, improving on content validity. Nevertheless, the WPT needed to have four-option items. The use section required four options to include the four parts of speech (noun, verb, adjective, and adverb). Having the same options of the four parts of speech for each item would be easier to work on than having three random parts of speech. The meaning section needed four options to decrease the effect of random guessing. As no *Don't know* options were provided, people who have no knowledge of an affix at all cannot help relying on random guessing. The length of options ranged between one and three words, which would not make a significant difference in time between three and four options. Similar to the meaning section, the form section needed four options to decrease the effect of random guessing. Each option had only one word, which would not seriously affect testing time.

5.5.2 Form

The form section aims to measure whether learners can recognise the written forms of affixes. After reviewing previous studies measuring knowledge of affix forms, this subsection discusses which test format is most appropriate for the WPT.

5.5.2.1 Previous Tests Measuring Affix Form

Very few attempts have been made to measure knowledge of affix forms. In an attempt to investigate the effects of Latin study on English vocabulary learning, Carroll (1940) developed a Morpheme Recognition Test, where learners must choose words that share

a common element of meaning from several words. Here is an example.

1. ready	
2. read	1. writing
3. regression	2. back, again
4. region	3. true
5. repeat	4. very
6. return	
7. rectangle	

For the example item, they must choose all the words in the left-hand column that have a common element of meaning. After that, they must choose the meaning of the word element from the four options in the right-hand column. The strength of this format is that it may decrease the effects of word knowledge. With a number of example words provided, each item does not rely on only one or two words to measure knowledge. However, this format requires knowledge of word meanings for finding a common element with a similar meaning, which may lower the construct validity. Another weakness may be that the format is complicated, because it may not be familiar to many learners.

A similar notion to recognition of affix forms is relational knowledge termed by Tyler and Nagy (1989). Relational knowledge has been measured in three ways. The most frequently used task is the "comes from" task first developed by Derwing (1970) (Berninger, Abbott, Nagy, & Carlisle, 2010; Costin, 1972; Kolstad, et al., 1985; Nagy, Berninger, Abbott, Vaughan, & Vermeulen, 2003; Nagy, Berninger, & Abbott, 2006). The task typically presents children with pairs of words, asking them to decide whether the pairs are derivationally related. For example, they are presented with the words *quick* and *quickness*, and asked to answer yes or no. An example of a "no" answer is the pair *moth* and *mother*.

Another test involves segmenting derived words into parts. Casalis and Louis-

Alexandre (2000) created a test which required children to segment a French affixed word into parts by pronouncing the parts separately. For example, they must pronounce *cassable* (=*breakable*) with a brief pause between *casse* and *able*.

Finally, Tyler and Nagy (1989) created a multiple-choice test where children must choose the meaning of an infrequent affixed word with a frequent base. Here is an example with the target word being *celebratory*.

> "I'm in a *celebratory* mood," Mary announced.
> Mary felt like:
> (a) having a party (b) being alone (c) going to sleep (d) having a fight (e) don't know

In this example, children need to recognise the relationship between *celebratory* and *celebrate*.

None of the three types of tests measuring relational knowledge is appropriate for the present research where the focus is on the recognition of affixes rather than word bases. Both the "comes from" task and Tyler and Nagy's (1989) task require knowledge of word bases. For example, it would be impossible for children who do not know the word *moth* to think of it as being unrelated to *mother* regardless of knowledge of the affix *-er*. It would also be impossible for those who do not know the word *celebrate* to demonstrate knowledge of the meaning of *celebratory*. Casalis and Louis-Alexandre's (2000) test may not require knowledge of affixes to get the items correct. For example, it would be possible to segment *breakable* into *break* and *-able* without knowledge of *-able* if the word *break* is known. There is a need to differentiate between knowledge of bases and knowledge of affixes.

5.5.2.2 Format for Form

There are three options for how to measure receptive knowledge of affix forms: (a) presenting nonsense words that include target affixes; (b) presenting real words that contain affixes; and (c) presenting affixes on their own without a base form.

The first option is to use nonsense words that contain affixes. For each item, test-takers are asked to recognize the affix that is found in the nonsense word. Here is an example for this format.

botodless (1) bo- (2) boto- (3) -less (4) -ss

In the example, test-takers must choose the affix in *botodless* from the four options. The first two options deal with the first few letters, and the last two options deal with the last few letters. The nonsense word was created from *colorless* by changing its consonants, based on Mochizuki and Aizawa (2000). The advantage of this format is that it uses nonsense words which ensure that the results are not affected by learners' prior knowledge of word bases. However, this format has a number of disadvantages that are difficult to overcome. A major prolem is that affixes with no recognisable bases are difficult to recognise. In the example above, *botodless* is supposed to be divided into *botod* plus *-less*. However, it would also be possible to divide it into *botodle* plus *-ess* which means 'female' such as *actress* and *princess*. Another problem is that materials development is difficult because creating nonsense words with real affixes is extremely difficult. It is almost impossible to invent phonologically, orthographically, morphologically, and etymologically correct nonsense words because each affix has a strong limitation to its base. For example, *gronersion* is not possible, because although there are words such as *gravitation*, *gr-* is more often Germanic (sometimes French), while *-ersion* is Latin. In addition, test-taking strategies may be used for this format,

which may lower the construct validity. A pilot study showed that in most cases test-takers arrived at correct answers without looking at the target nonsense words. In the example above, they only looked at the four options, thinking that among them *-less* was most likely to be an affix so it should be an answer. This may indicate that nonsense words are not necessary. For these reasons, this format is not used for the WPT.

The second option is to use real target words containing affixes and segment real words into an affix and a base. Here is an example.

endless (1) en-dless (2) end-less (3) endl-ess (4) endle-ss

In the example, test-takers must choose the answer that correctly divides *endless* into its base and its affix. The target word *endless* was the most frequent word with the suffix *-less* in the BNC.[28] This format resolves some problems with the format using nonsense words as target words. First, the use of real words ensures that affixes are recognisable because the bases are real and recognisable. Second, materials development is not difficult because there is no need to create nonsense words. However, it is not clear what is being measured in this format; that is, it measures knowledge of base forms as well as affix forms. Knowledge of word bases may be sufficient for choosing correct answers. For example, a pilot study revealed that some low-proficiency learners got the example item correct by thinking like this: "I don't know *endless* or *-less*, but can see the familiar word *end*. I don't know the words in the other options. So *end-less* should be the answer." For this reason, this format was considered inappropriate for the WPT.

The last option is to avoid using example words and present affixes in isolation.

[28] In the same way as the first format with nonsense target words, the four options could have involved only affixes such as (1) en-, (2) end-, (3) -less, and (4) -ess. This format may be too complicated, because each option could be an affix without the target word *endless*.

Here is an example.

(1) -ique (2) -less (3) -eeve (4) -itle

Test-takers must choose an affix from four options with the same number of letters. The three distractors are a string of letters that appear in English but are not affixes. In the example, *-ique* occurs in words such as *technique*, *-eeve* in *sleeve*, and *-itle* in *subtitle*; however, they are not English suffixes. In this format, knowledge of English orthography and phonology is not sufficient for eliminating distractors because all the options are orthographically and phonologically real in English. This format has advantages over the two formats discussed above. First, no use of target words ensures that the results are not affected by their prior knowledge of word bases. Second, materials development is not difficult because there is no need to create nonsense words. The potential problem with this format is that it may underestimate knowledge of affix forms because test-takers need to recall some example words beginning or ending with particular letters unless they have explicit knowledge of affix forms. In order to examine whether those who have knowledge of an affix form can demonstrate their knowledge, a pilot study was conducted with two native speakers teaching university-level English and five proficient non-native speakers doing their Ph.D. in applied linguistics. The results showed that while they had no difficulty finding answers for prefixes, they found it difficult to determine the answers for some suffixes, especially infrequent short suffixes such as *-i* in *Israeli* and *-et* in *owlet*. This indicates that although this format may underestimate knowledge of some suffix forms, it works well for prefixes and the majority of suffixes. In conclusion, this format is considered to be most appropriate for the form section.

5.5.2.3 Target Affixes

All the affixes are included in the form section, but affixes that have the same written form are treated as one item. For example, there are two types of *-al* (making a noun and making an adjective), but there is no differentiating between the two affixes on this format; thus, these two affixes are presented as one item. This section has a total of 107 items.

5.5.3 Meaning

The meaning section aims to measure whether learners can demonstrate knowledge of written receptive affix meanings. After reviewing previous studies measuring knowledge of affix meanings, this subsection discusses which test format is most appropriate for the WPT.

5.5.3.1 Previous Tests Measuring Affix Meaning

Very few attempts have been made to measure learners' knowledge of affix meanings. One study was done by Mochizuki (1998) who investigated Japanese learners' English affix knowledge. Each item on the test had three example words with the affix which were supposed to be familiar to learners underlined. They were required to choose one correct meaning from a set of four options written in Japanese. The options were created based on a pilot study in which five university students were asked to productively provide the meaning of each affix. Here is an example.

<u>auto</u>graph	<u>auto</u>genous	<u>auto</u>nomy
(1) 自らの (self)		(2) 真の (real)
(3) 自動の (automatic)		(4) 人工の (artificial)

There are three major problems in Mochizuki's format. First, he focused on prefix meanings and did not measure knowledge of suffix meanings. It seems that some affixes such as *-able* and *-less* add clearer lexical meanings to their bases than others such as *-ive* and *-ness*. For example, *endless* is an adjective that means 'without end', where *-less* adds the meaning of 'without' to its base *end* as well as changes the part of speech of *end* from a noun to an adjective. The meaning of *-less* is worth measuring, because knowing that *-less* makes an adjective is not sufficient to understand the meaning of *endless*. In fact, *endless* is not simply an adjective version of *end* which means 'relating to *end*'.

Second, clear criteria for selecting example words were not given. Although Mochizuki admitted that the results might have been affected by the words presented as examples, he did not give explicit description of the ways in which he selected the example words. He chose supposedly familiar words to the learners, but it is unclear how he determined familiarity. The three words in the example above are not frequent: *autograph* and *autonomy* belong to the 9,000-word level and *autogenous* is beyond the 10,000-word level in the BNC word lists. While *autonomy* appears 1,814 times in the BNC, *autograph* appears 174 times and *autogenous* appears only twice.

Finally, some options overlapped in meaning, which might have made the test unreliable. Mochizuki reported that the learners may have confused Options 1 (self) and 3 (automatic), because the two options were semantically similar.

Another study by Mochizuki and Aizawa (2000) attempted to overcome Mochizuki's (1998) flaws by using pseudowords as target words. For each item, test-takers were presented with three pseudowords created by changing the consonants of real words. They were asked to choose one correct meaning from a set of four options

written in Japanese, their first language. Here is an example.

exlorp exckanze exnanx (1) following (2) causing (3) out (4) including

The major problem with this format is that success may depend on which words test-takers recall for each affix. For example, the correct answer to the example above should be Option 3 (out). However, if they recall words such as *ex-wife* and *ex-president*, they might choose Option 1 *following* which indicates a sequential order of time. If they recall words such as *exact* and *example*, it is almost impossible to choose an answer, because *exact* does not mean '*out* act' and *example* does not mean '*out* ample'. Without the word base, it is difficult to determine the meaning.

5.5.3.2 Format for Meaning

There are three possible options for how to present the affixes: (a) presenting affixes in isolation; (b) presenting affixes with nonsense words; and (c) presenting affixes with real words.

The first option for target affixes is to present them in isolation. Here are two examples.

-less (1) before (2) without (3) the furthest (4) person de- (1) opposite (2) person/thing (3) together (4) small

In this format, test-takers must choose the meaning of the affix for each item. Distractors are the meanings of other randomly chosen affixes. The advantage of this format is that the results are not affected by prior knowledge of word bases because no

real words are presented. However, this format has two major problems. The first problem is that this format may measure productive rather than receptive knowledge, because test-takers need to produce example words with the affixes. For the example *-less* above, they need to recall words with the *-less* ending such as *endless* and *useless* unless they have explicit knowledge of *-less*. In order to examine whether people with knowledge of an affix can demonstrate their knowledge of its meaning, a pilot study was conducted with five native speakers studying applied linguistics for their MA. It indicated that it was difficult to recall words with some affixes such as *de-* because some participants confused *de-* and *di-*. With some example words provided, they had no difficulty finding the correct answer. It should be reasonable to assume that L2 learners may not be able to demonstrate their knowledge of affix meanings due to inability to recall appropriate words. The other problem is that success depends on words that are recalled. For the example *de-* above, if words such as *define* and *detail* are recalled, it may be difficult to choose the meaning of *de-* (opposite), because *define* does not mean 'the opposite of *fine*' and *detail* does not mean 'the opposite of *tail*'. It follows that wrong responses do not necessarily mean lack of knowledge of affix meanings; that is, they may be due to (1) ignorance of the affix, (2) inability to recall words containing the affix despite knowledge of it, or (3) recalling words containing the affix that do not convey the target meaning. Thus, the construct being measured in this format is unclear due to the ambiguity of the interpretation of wrong responses. For these reasons, this format was not used for the present research.

The second option for target affixes is to present them in nonsense words. Here are two examples.

botod<u>less</u>

(1) before (2) without (3) the furthest (4) person

<u>de</u>goze

(1) opposite (2) person/thing (3) together (4) small

In this format, each item has a nonsense word base with a real affix which is underlined so that the affix in question may be easily recognised. Test-takers must choose the meaning of the underlined affix. Distractors are the meanings of other randomly chosen affixes. This format has the advantage of being independent of prior knowledge of word bases because no real words are presented. However, it has the same problems as the above-mentioned format (presenting affixes in isolation); that is, it may measure productive knowledge because test-takers need to recall example words with the target affixes, and success depends on which example words are recalled. Thus, this format was considered inappropriate for the WPT.

The last option is to use real words that contain the affixes. Here are two examples.

-less (end<u>less</u>; use<u>less</u>)

(1) before (2) without (3) the furthest (4) person

de- (<u>de</u>compose; <u>de</u>code)

(1) opposite (2) person/thing (3) together (4) small

For each item, a target affix is followed by two example words with the affix underlined for easy recognition. Test-takers must choose the meaning of the affix represented in the two example words. The instructions state that the affix can attach to other words than the two example words. Similar to the previous formats, distractors are the meanings of other randomly chosen affixes. Two example words are provided in case one is unknown. This is possible because each affix appears in at least two word families in the top 10,000 word families in the BNC word lists. This format may be a solution to

the problems that the other two formats have. Test-takers do not need to recall example words with the target affixes because two example words are presented for each affix. However, the weakness of this format is that success may depend on knowledge of example words given. In order to overcome this weakness, example words are selected from words that meet the following three criteria: frequency, semantic transparency, and regularity in connection. The first criterion is frequency. The example words are chosen from highly frequent words in order to maximise the likelihood that test-takers will know the example words. The word frequencies are calculated based on the BNC. For example, *endless* and *useless* are chosen as the example words for the affix *-less* because these words are the most frequent words with *-less* in the BNC.[29] Second, the example words are semantically transparent in order to maximise the likelihood that test-takers will be able to demonstrate their receptive knowledge of affix meanings even if they have no explicit knowledge of the affix meanings. Semantic transparency is estimated based on Nagy and Anderson's (1984) six levels of semantic relatedness. The level refers to the degree to which a derivational word is inferable from its base. A brief description of each level is given in Table 42, with the upper levels being more semantically transparent and the lower levels being less semantically transparent.

[29] To be precise, the top ten words with the *-less* ending are *unless* (10,838 times of occurrence), *nevertheless* (7,045), *endless* (1,532), *regardless* (1,532), *nonetheless* (1,296) *useless* (1,261), *homeless* (1,065), *doubtless* (844), *helpless* (792), and *hopeless* (712). *Unless, nevertheless,* and *nonetheless* are not affixed words. In most cases (1,396 out of 1,532 times of occurrence), *regardless* is used as part of the phrase *regardless of* which means 'in spite of' and is not semantically clear. Thus, *endless* and *useless* are selected for the example words.

Table 42. Degrees of semantic relatedness

Level	Description	Example
0	The meaning of the derivative can be inferred from the meaning of its base	senselessly; senseless
1	The meaning of the derivative can be inferred from the meaning of its base with minimal help from context	various; vary
2	The meaning of the derivative can be inferred from the meaning of its base with reasonable help from context	theorist; theory
3	The meaning of the derivative includes semantic features that are not inferable from the meaning of its base without substantial help from context	visualise; visual
4	The meaning of the derivative is related to the meaning of its base, but only distantly	saucer; sauce
5	There is no discernible semantic connection	prefix; fix

The example words are selected from words at Level 0 or 1 in Nagy and Anderson's classification of semantic relatedness. Finally, each example word has the target affix and its base that are regularly connected in order to maximise the likelihood that test-takers will recognise both the target affix and its base without difficulty. Regularity in connection refers to the degree of change in spelling when an affix is added to its base. For example, *discussion* is regular in connection because it is made from *discuss* and *-ion* without any unpredictable change in the base. *Permission*, on the other hand, is not regular in connection, because it is made from *permit* and *-ion* with *t* in *permit* changed into *ss*. Only example words that are regular in connection are used. In summary, this format is weak in that success may depend on knowledge of the example words given, but this weakness may be minimised by selecting example words that meet the three criteria (frequency, semantic transparency, and regularity in connection). In conclusion, this format was considered to be most appropriate for the meaning section.

5.5.3.3 Target Affixes

For the meaning section, affixes that have abstract meanings such as *-ness* (state, condition, quality) in *happiness* and *-ment* (action, state, results) in *movement* were excluded. A pilot study was conducted with ten native speakers doing their Ph.D. or MA in linguistics or applied linguistics to see whether they could specify the meanings of these affixes. For each affix, they were presented with two words that contained the affix and were asked to write the meaning of the affix. A total of 73 affixes whose meanings could be explicitly described by eight or more native speakers were selected for this section.

The meaning of each affix was largely based on the Oxford Advanced Learner's Dictionary (7th edition). If an affix had multiple meanings, the most frequent meaning was measured; for example, "in advance" (e.g., *foresee* and *forewarn*) was chosen for the meaning of *fore-* instead of "in front of" (e.g., *forehead* and *forearm*) because the affixes with the former meaning appeared more frequently than the latter in the first 10,000 word families in the BNC word lists. Meanings were briefly paraphrased with high-frequency words (most frequent 1,000 words) so that even low-proficiency learners may understand.

5.5.4 Use

Some affixes have the function of changing the part of speech of the word base. For example, *-ment* attaches to verbs such as *move* and *develop*, changing their part of speech to nouns (*movement* and *development* are nouns). The use section aims to measure whether learners can demonstrate knowledge of the part of speech that an affix makes. After reviewing previous studies measuring knowledge of affix functions, this

subsection discusses which test format is most appropriate for a standardised test of affix function.

5.5.4.1 Previous Tests Measuring Affix Use

There are a number of studies that investigated L1 children's knowledge of the syntactic properties of affixes. The tasks measuring it are broadly classified into three types: definition, sentence completion, and judgement tasks.

A definition task measures knowledge of affix functions by getting children to write the definition of the target word containing an affix (Wysocki & Jenkins, 1987). For example, children are asked to write the meaning of the word *existence* which contains the target suffix *-ence*. The answer *living* was credited as correct because both the meaning and the syntax were correct, while the answer *alive* was scored as only partially correct because the meaning was correct but the syntax was not correct. This format is not appropriate for the WPT, because it requires the ability to write the definition in English, which may be demanding for L2 learners with a low level of proficiency. It also takes time to complete and grade.

A completion task asks children to choose a word with a correct derivational suffix that fits into the blank in a sentence. This task was first developed by Tyler and Nagy (1989) and subsequent studies used their format to measure knowledge of suffix functions (Costin, 1970; Mahony, 1994; Nagy, et al., 2003; Nagy, et al., 2006). In an attempt to investigate the acquisition of English derivational morphology by children in the fourth, sixth, and eighth grades, Tyler and Nagy (1989) selected 16 derivational suffixes, each of which was measured twice using a real target word and a nonsense

target word. For each item, the children must choose the correct derivative that fits into the blank in a sentence. Here is an example of real-word items.

You can ___ the effect by turning off the lights. (1) intensify (2) intensification (3) intensity (4) intensive

This format is strong in authenticity because it avoids the need for metacognitive knowledge about parts of speech, but the construct being measured may not be clear. Knowledge of affix functions is not sufficient for arriving at the correct answers, because the syntactic property of the blank needs to be identified from the context.

Here is an example using nonsense-word items. For each item, the children must choose the nonsense word with the suffix that fits into the blank in a sentence.

I wish Dr. Who would just ___ and get it over with. (1) transumpation (2) transumpative (3) transumpate (4) transumpatic

In addition to the problem mentioned above, this format has another potential weakness. The answer should be Option 3, because the blank should be a verb and *-ate* occurs in verbs such as *activate* and *originate*. However, there are a number of nouns (e.g., *candidate* and *certificate*) and adjectives (e.g., *fortunate* and *passionate*) with the *-ate* ending. The part of speech of words with *-ate* cannot be controlled for by nonsense words. Option 2 might be a popular distractor, because there are verbs with the *-ive* ending such as *receive* and *arrive*. The results showed that the nonsense-word format was much more difficult than the real-word format.

Another type of the sentence completion task is to ask children to write the word with an appropriate suffix that fit into the blank in a sentence (Berninger, et al., 2010; Kolstad, Kolstad, & Wagner, 1986). For example, children are presented with the word *farm*, and are asked to fit it into the blank in the sentence *The ____ is plowing his fields*.

This format is not appropriate for the WPT because it requires the ability to comprehend the context and knowledge of the base as well as the affix.

Finally, a judgement task typically asks children to choose one sentence in which a word with a derivational suffix is correctly used. Tyler and Nagy (1990) asked children to choose the most appropriate paraphrase of the top sentence from four options. Here is an example.

> Mary was afraid that a general indecision about the use of nuclear weapons might be a threat to national security.
>
> (a) Mary feared that if most people couldn't make up their minds about using atomic bombs, the country could be put in danger.
> (b) Mary feared that a military officer who couldn't make up his mind about using atomic bombs might put the country in danger.
> (c) Mary feared that a public discussion about using atomic bombs might put the country in danger.
> (d) Mary feared that a military officer who openly discussed using atomic bombs might put the country in danger.

In this example, the target word was *indecision* and Option (a) was the correct answer. In Option (b), the meaning of *indecision* was correct, but its syntax was wrong: *Indecision* was mistaken as *indecisive* which modified the noun *general*. In Option (c), the syntax of *indecision* was correctly interpreted as a noun, but its meaning was not correct. Option (d) was wrong in both the meaning and the syntax of *indecision*. This format is not appropriate for the WPT because it is demanding and time-consuming. Test-takers need to read 98 running words for this item. They also need to recognise that *indecision* is paraphrased in a variety of ways and is the key to finding the correct answer. This format may measure the ability to comprehend the sentences as well as knowledge of affix functions.

In an attempt to improve the shortcomings in Tyler and Nagy's (1989) measurement method, Nagy, Diakidoy, and Anderson (1993) selected 20 rarely

occurring suffixed words with frequent bases and asked fourth-grade, seventh-grade and high-school students to choose sentences in which rarely suffixed words are used correctly. (A similar format was also used by Nagy, et al. (2006).) Here is an example. The students must choose a sentence in which *powderize* is correctly used.

(a) First they had to find a powderize rock.
(b) First they had to powderize find the rock.
(c) First they had to find a powderize for the rock.
(d) First they had to find a way to powderize the rock.

The strength of this format is that it can avoid the need for metacognitive knowledge of part of speech. It is also similar to authentic situations where affix knowledge is used for inferring the meaning of an unknown word with a familiar base. However, the use of context requires the comprehension of the context as well as knowledge of affixes. In the example, the students need to understand the syntactic property of *powderize* from a given context in each option. Option (d) is correct, but *to* which is followed by *powderize* has to be interpreted as an infinitive marker instead of a preposition. In Option (b), *powderize* also follows *to*, but the students have to recognise that *powderize* is used as an adverb. In addition, this format is time-consuming: the students need to read 37 words in total for one item. The predominance of grammatically incorrect sentences may also have a negative impact on learning..

In summary, L1 children's knowledge of affix functions has been measured in context to avoid metacognitive knowledge of parts of speech. The use of context essentially involves comprehension of the context as well as knowledge of affix functions in the construct being measured. It also takes time to work on the items and is not appropriate for the WPT which needs to be easily completed and graded.

More explicit knowledge of affix functions has been measured for L2 learners. The first empirical study with L2 learners was done by Schmitt and Meara (1997). In their

receptive format, learners must choose all the suffixes that can be attached to each target verb from a set of 14 suffixes. Here is an example with the target verb *use*.

use -able -age -al -ance/ence -ed -ee -er/or -ing -ion -ive -ly -ment -s -ure

This format seems to measure distributional knowledge (knowledge of which classes of stems can take certain affixes). The WPT does not aim to measure this type of knowledge.

Mochizuki (1998) created a test that explicitly measured knowledge of affix functions. Each item had three real words that were supposed to be unfamiliar to the test-takers, followed by the four options: noun, verb, adjective, and adverb. Here is an example with the target suffix *-al*.

dismissal reversal avowal 名(noun) 動(verb) 形(adjective) 副(adverb)

Although real words are presented, they are supposed to be unfamiliar to the test-takers in order to avoid the use of knowledge of the example words instead of affix knowledge. Mochizuki later noted that he was unable to control for familiarity and the words were not always unfamiliar to the test-takers.

Mochizuki and Aizawa (2000) preferred to use nonsense words to real words to make sure that the example words were unfamiliar to all the test-takers. Here is an example with the target affix *-al*.

dutical ravional nolisical (1) noun (2) verb (3) adjective (4) adverb

Three nonsense words are presented with the target affix underlined. In the example, although *-al* could make a noun, Option (3) *adjective* was the only correct answer

because adjectival examples are more frequent than noun examples. The weakness of this format is that it cannot deal with affixes with multiple functions.

In summary, a review of the literature that measured knowledge of affix functions showed that none of the test formats would be appropriate for the WPT.

5.5.4.2 Format for Use

This subsection first discusses whether context should be provided for the use section, and then examines how to present affixes based on Bachman and Palmer's (1996) test usefulness.

A decision has to be made as to whether to use context to measure knowledge of affix functions. Contextualised formats, which measure affix knowledge in context, have been typically used to measure L1 children's knowledge of affix function, because it can avoid the need for metacognitive knowledge of part of speech (e.g., Nagy, et al., 1993; Tyler & Nagy, 1989). However, as mentioned earlier, affix knowledge and comprehension of context are confounded in this format. It also takes time to complete because test-takers need to read a number of words in the context. On the other hand, decontextualised formats, which measure explicit knowledge of affix function without using context, have been used to measure L2 learners' affix knowledge (Mochizuki, 1998; Mochizuki & Aizawa, 2000). This format is easy to complete because test-takers have only to choose the part of speech that an affix makes from a set of options. The construct definition is clear because this format measures whether test-takers know the syntactic property that the affix has. The weakness is that it requires metacognitive knowledge of part of speech and may underestimate knowledge of affixes. However, no problems were reported in Mochizuki (1998) and Mochizuki and Aizawa (2000)

concerning the use of decontextualised formats, suggesting that L2 adult learners may have explicit knowledge of part of speech. For the present research, decontextualised formats are considered appropriate because the aim is to create a word part test that measures adult L2 learners' overall affix knowledge and can be easily completed and graded. The weakness of the need for metacognitive knowledge of part of speech may be minimised by providing an example sentence for each part of speech at the beginning of the use section.

There are three possible options for how to present the target affixes: (a) presenting affixes in isolation; (b) presenting affixes with nonsense words; and (c) presenting affixes with real words.

The first option for target affixes is to present affixes in isolation. For each item, the target affix is presented on its own, and test-takers must choose its syntactic property from the following four options: noun, verb, adjective, and adverb. Here are two examples.

-less
(1) noun (2) verb (3) adjective (4) adverb
-al
(1) noun (2) verb (3) adjective (4) adverb

This format has the advantage of being independent of prior knowledge of word bases because no real words are presented. However, it has two major problems. First, as with the first option for the meaning format, it may measure productive knowledge rather than receptive knowledge because test-takers may need to recall one or two example words. The other problem is that the scoring of affixes with multiple functions is difficult. For the example *-al* above, both Options 1 (noun) and 3 (adjective) are correct, because *-al* has two functions (making a noun and an adjective). The solutions to this

problem are (1) giving credit only to the more (or the most) frequent part of speech, (2) allowing multiple correct answers, and (3) omitting the other possible answers from the options. The first solution, which was used by Mochizuki and Aizawa (2000), regards an adjective as only the correct answer, because adjectives with the *-al* ending appear more frequently than nouns with the *-al* ending. This scoring method cannot tell whether people who chose a noun knew that *-al* had the function of making an adjective. Second, a multiple-choice format with multiple correct answers is demanding because test-takers need to examine each option carefully. For example, they need to recall nouns containing *-al* such as *approval* and *proposal* to conclude that *-al* has the function of making a noun. They also need to check whether words with the *-al* ending can be verbs and may recall words such as *reveal* and *signal* which end with *-al* but do not have the affix *-al*. Similar processes are needed for adjectives and adverbs. The final solution is to have only one correct answer and three distractors. For *-al*, the options could be (1) preposition, (2) verb, (3) adjective, and (4) adverb in order to allow only one correct answer ((3) adjective). This format cannot measure whether learners know that the affix *-al* makes a noun. In summary, none of the three solutions is effective for measuring written receptive knowledge of affix function.

The second option is to present nonsense words containing real affixes. For each item, test-takers must choose its syntactic property from the following four options: noun, verb, adjective, and adverb. Here are two examples.

botod<u>less</u>
(1) noun (2) verb (3) adjective (4) adverb
verton<u>al</u>
(1) noun (2) verb (3) adjective (4) adverb

In this format, each item has a nonsense word base with a real affix which is underlined so that test-takers can recognise the affix in question. This format has the advantage of being independent of prior knowledge of word bases because no real words are presented. However, it has the same problems as the above-mentioned format (presenting affixes in isolation); that is, it may measure productive knowledge because test-takers need to recall example words with the target affixes, and the scoring is difficult because the answers for affixes with multiple functions cannot be controlled for. Thus, this format was considered inappropriate for the WPT.

The last option is to present affixes in real words. Here are three examples.

-less (end<u>less</u>; use<u>less</u>)
(1) noun (2) verb (3) adjective (4) adverb
-al (person<u>al</u>; tradition<u>al</u>)
(1) noun (2) verb (3) adjective (4) adverb
-al (propos<u>al</u>; approv<u>al</u>)
(1) noun (2) verb (3) adjective (4) adverb

For each item, a target affix is followed by two example words with the affix underlined so that test-takers can recognise it. The instructions state that the affix can attach to other words than the two example words. Two example words are provided in case one is unknown. The example words aim to help test-takers to recognise the function of the affix and to control for the part of speech the affix makes. As with the meaning format, the example words are chosen from words that meet the following three criteria: frequency, semantic transparency, and regularity in connection. For affixes with multiple functions, example words are carefully chosen so that both example words will have the same function. For example, *-al* has the functions of making both adjectives and nouns. As illustrated in the example above, *-al* in *personal* and *traditional* makes an

adjective (and not a noun), whereas *-al* in *proposal* and *approval* makes a noun (and not an adjective). The following eight affixes have multiple functions: *-al* (adjective in *personal* and noun in *proposal*), *-ant* (adjective in *pleasant* and noun in *consultant*), *-ary* (adjective in *revolutionary* and noun in *secretary*), *-ate* (adjective in *fortunate* and verb in *activate*), *-en* (adjective in *wooden* and verb in *weaken*), *-ent* (adjective in *different* and noun in *respondent*), *-ly* (adjective in *friendly* and adverb in *clearly*), and *-y* (adjective in *lucky* and noun in *difficulty*). This format may be a solution to the problems that the other two formats have. Test-takers do not need to recall example words with the target affixes because two example words are presented for each affix. In addition, the correct answer can be controlled for by presenting two example words. However, the weakness of this format is that success may depend on knowledge of example words given. As with the meaning format, this weakness may be minimised by providing example words that are most frequent, semantically transparent, and regular in connection. Another potential weakness is that test-takers may find an answer based on knowledge of the example words without attributing the grammatical function to the affix. For example, they may easily find an answer for the item *-less* without any help from the suffix if they know that *endless* is an adjective in the same way as knowing that *happy* is an adjective. However, the WPT can examine whether or not test-takers recognise *-less* as an analysable part in the form section. If they can recognise *-less* as an analysable part and think of *endless* as an adjective, then it should be reasonable to assume that they know *-less* has the function of forming an adjective either implicitly or explicitly because the base *end* is not an adjective. In conclusion, this format was considered most appropriate for the WPT.

5.5.4.3 Target Affixes

For this section, 59 class-changing affixes (four prefixes and 55 suffixes) were included. The four prefixes were *a-* (aside), *be-* (belittle), *em-* (empower), and *en-* (enrich). Some suffixes are class-maintaining (no change in the part of speech) and were not included. For example, *-ette* attaches to nouns such as *kitchen* and *cigar*, resulting in nouns such as *kitchenette* and *cigarette*. The 13 suffixes of this type were *-dom* (kingdom), *-eer* (mountaineer), *-ess* (princess), *-et* (owlet), *-ette* (kitchenette), *-ful* (handful), *-hood* (childhood), *-ism* (Darwinism), *-ist* (artist), *-let* (booklet), *-ling* (duckling), *-ship* (friendship), and *-ster* (gangster).[30] The score interpretation of these class-maintaining suffixes is difficult because correct responses may be due to either knowledge of the base or knowledge of the suffix. Suppose that a test-taker knows the word *duck* but does not know the word *duckling* or the affix *-ling*. He or she may be able to get this item correct by thinking that *duckling* must be related to *duck* which is a noun, so *duckling* is most likely to be a noun. Other suffixes that were not included in the use section were *-an* (American), *-ese* (Japanese), *-fold* (twofold), *-i* (Israeli), *-ian* (Egyptian), *-ite* (Israelite), *-most* (topmost), and *-th* (fourth). These suffixes were excluded because they have multiple functions that cannot be controlled for by presenting example words. For example, *-an* in *American* and *European* has the functions of making both a noun and an adjective.

In order to make sure that each item had only one correct answer, a pilot study was conducted where ten PhD candidates (five native and five non-native speakers) studying linguistics or applied linguistics individually answered the items of the use section. The

[30] It should be noted that some of the suffixes change the part of speech of the bases. For example, *-dom* attaches to adjectives such as *free* and *wise*, resulting in nouns such as *freedom* and *wisdom*. However, many more words with the suffix *-dom* in the most frequent 10,000 word families maintain the part of speech of the bases than those that change the part of speech of the bases (e.g., *kingdom, stardom, Christendom, earldom,* and *dukedom*).

results showed that nine or ten of the participants answered each item correctly, indicating that the example words were appropriate for determining the part of speech formed by the 59 target affixes including the ones with multiple functions (e.g., *-al*).

5.6 Summary

For the purpose of selecting useful affixes for vocabulary learning, the present research has identified a total of 118 affixes (42 prefixes and 76 suffixes) based on the following three criteria: (1) They are bound morphs which attach to free morphs; (2) They appear in more than one word family in the most frequent 10,000 word families in the BNC word lists; and (3) Allomorphs are treated as different affixes. The selected suffixes covered the majority of the suffixes used in previous studies. The excluded suffixes included inflectional suffixes (e.g., *-ed* and *-s*) and semi-suffixes (e.g., *-like*). The coverage of the selected prefixes to those identified in earlier studies was not very high, mainly because prefixes that do not attach to free morphs (e.g., *ad-* and *com-*) were included in previous studies.

The literature (Bauer & Nation, 1993; Nation, 2001; Tyler & Nagy, 1989) indicated that written receptive knowledge of affixes involved three aspects: form (recognising written forms of affixes), meaning (recognising meanings of affixes), and use (recognising syntactic properties that affixes have). These three aspects of affix knowledge were measured in three different sections: form, meaning, and use sections. A sample format for these sections is presented in Table 43.

Table 43. Test format for the word part test (an example for *-less*)

Section	No. of items	Format
Form	107	(1) -ique (2) -less (3) -eeve (4) -itle
Meaning	73	-less (endless; useless) (1) before (2) without (3) the furthest (4) person
Use	59	-less (endless; useless) (1) noun (2) verb (3) adjective (4) adverb

The form section asks test-takers to choose the affix form from four options. The three distractors are real strings of letters in English but are not affixes which change the meaning or the syntactic property of the bases. The meaning section requires test-takers to choose the meaning of the target affix from four options. The three distractors were randomly chosen from the meanings of other affixes. For the use section, test-takers need to choose the part of speech that the affix makes from the following four options: noun, verb, adjective, and adverb. For the meaning and the use sections, two example words are presented for each item to help test-takers demonstrate their knowledge of affix meanings and functions even if one is unknown. Providing two of them is possible because the affixes appear in at least two words in the most frequent 10,000 word families. The example words are the most frequent, semantically transparent, and regularly connected words to maximise the likelihood that they would know the example words.

CHAPTER 6

VALIDATION OF THE WORD PART TEST

This chapter describes the validation of the word part test (WPT). Poorly written items were identified and rewritten based on Rasch analysis (Study 1). The WPT was revised based on the results in Study 1, and the validity of the revised WPT was discussed from eight aspects of construct validity (Study 2). This chapter also discusses theoretical values of the WPT and provides a proposal for score interpretation and reporting to learners.

6.1 Study 1

The purpose of Study 1 was to identify and rewrite poorly written items based on Rasch analysis. Poor items needed to be rewritten instead of being simply omitted so that all the affixes that were selected for the present research would be included in the WPT.

6.1.1 Participants

A total of 417 Japanese university students (273 males, 136 females, and 8 unspecified) learning English as a foreign language participated in the research.[31] The WPT was administered to university students from 19 intact classes at six different universities (see Table 44). The participants' ages ranged between 18 and 23 with the average being 19.1 (SD=5.0). The participants had had at least six years of prior English instruction

[31] Although a total of 440 participants took the test, the data from 417 participants were analysed. The 23 participants who were excluded from the analysis included those who left latter items unanswered, or gave answers without thinking seriously (e.g., choosing Option 2 for every item).

Table 44. Description of participant groups

University	No. of classes	*N*	Purpose of English Education
A	5	119	English for Academic Purposes
B	2	56	English for General Purposes
C	2	55	English for General Purposes
D	5	69	English for General Purposes
E	3	69	English for Business Purposes
F	2	49	English for Business Purposes
Total	19	417	

(three years at junior-high and three years at senior-high school). Their majors included agriculture, economics, engineering, law, literature, medicine, and pharmacology.

The participants varied widely in their English proficiency levels. The self-reported TOEIC scores from 67 of the participants were summarised as follows: Mean=509.0, SD=141.6, Max=880, Min=235.[32] The distribution is illustrated in Figure 28, indicating a wide range of proficiency levels of the participants.

The participants' English vocabulary sizes were also estimated through a bilingualised Japanese-English version of Nation and Beglar's (2007) Vocabulary Size Test (available at http://www.victoria.ac.nz/lals/staff/paul-nation.aspx). The Vocabulary Size Test was administered to 238 participants (57.1% of the participants who took the WPT) and their estimated vocabulary sizes ranged between 4,000 and 10,100 word families (Mean=7,290, SD=954). The distribution is illustrated in Figure 29.

[32] The TOEIC scores available may not be fully representative of the participants, because students in some classes were required to take TOEIC, whereas others were not. However, the purpose here is to show that the test was administered to learners with a wide range of proficiency.

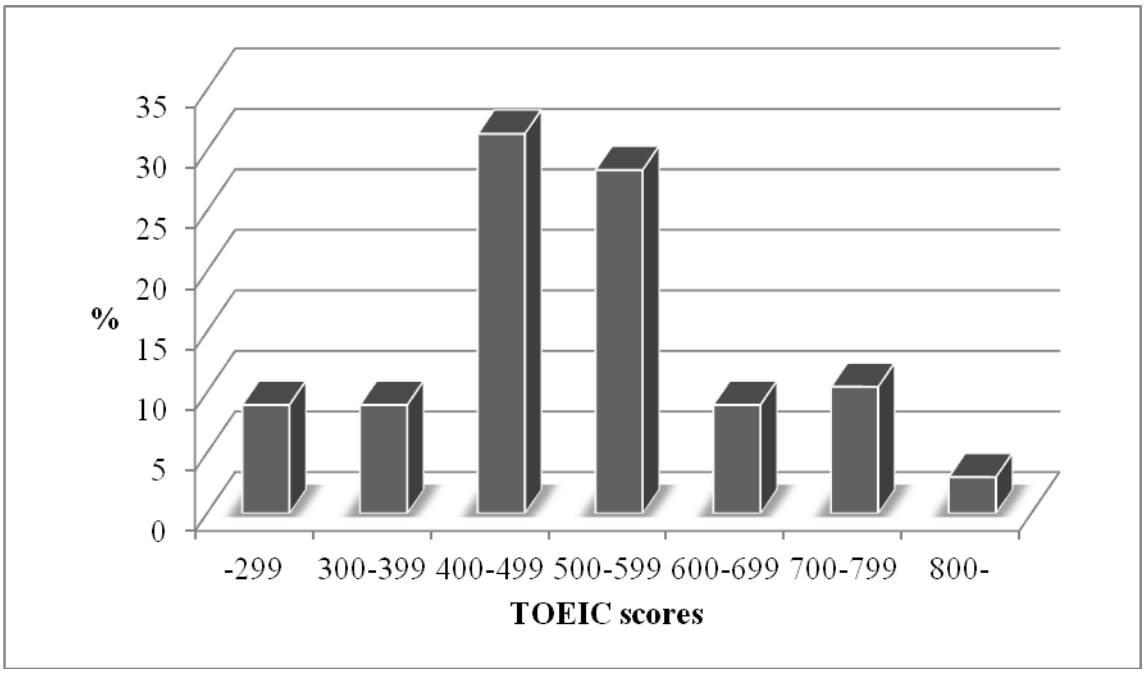


Figure 28. Proficiency range (TOEIC scores)

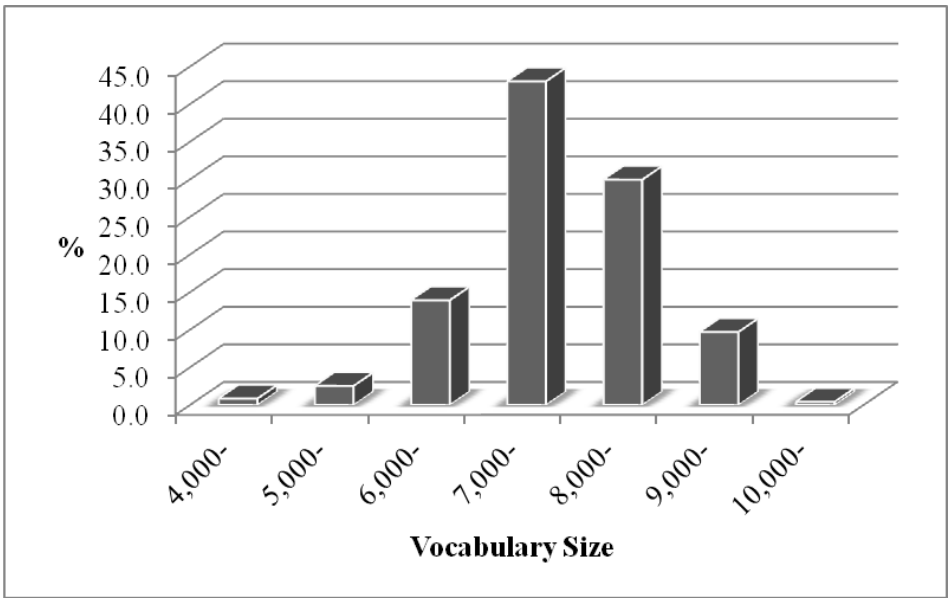


Figure 29. Vocabulary size range

6.1.2 Materials

As discussed in the previous chapter, the WPT had three sections: form, meaning, and use. For the form section, test-takers must choose the real affix from four options with

three distractors. Here are two examples from this section.

1. (1) non- (2) kno- (3) spo- (4) orn-
2. (1) -rse (2) -ack (3) -ful (4) -uin

For the meaning section, test-takers must choose the closest meaning of the target affix. For each item, two example words containing the target affix were provided. Here are two examples from this section.

1. dis- (<u>dis</u>believe; <u>dis</u>similar)
   (1) Not
   (2) Person
   (3) New
   (4) Main
2. -ist (special<u>ist</u>; art<u>ist</u>)
   (1) against
   (2) person
   (3) two
   (4) not

For the use section, test-takers must choose the part of speech that the target affix makes. For each item, two example words containing the target affix were provided. Each item had a fixed set of four options: noun, verb, adjective, and adverb which were translated into Japanese (名詞, 動詞, 形容詞, and 副詞) because it was predicted that the majority of low-proficiency learners would be unfamiliar with the four parts of speech in English. Here are two examples of this section.

1. en- (<u>en</u>danger; <u>en</u>large)
   (1) 名詞
   (2) 動詞
   (3) 形容詞
   (4) 副詞
2. -ful (care<u>ful</u>; use<u>ful</u>)
   (1) 名詞
   (2) 動詞
   (3) 形容詞
   (4) 副詞

The test length for Study 1 was determined so that the test scores would be most reliable. As a general principle, increasing the number of items leads to better reliability (e.g., Bachman, 1990), but having too many items may decrease reliability due to a fatigue effect. In order to examine the number of items that would maximise the reliability of the WPT, a pilot study was conducted with six Japanese learners of English at a beginner level. They individually took the WPT with three different lengths. The

first two participants took the whole test (107 items for the form section, 73 items for the meaning section, and 59 items for the use section). The results showed that they got tired of the test towards the end of it and gave up completing the test. Two of the other participants took three quarters of the items (80 items for the form section, 55 items for the meaning section, and 44 items for the use section). They could complete the test, but a post-test interview revealed that they felt the test was too long. Thus, a slightly shorter version of the WPT with two thirds of the items (75 items for the form section, 49 items for the meaning section, and 40 items for the use section) was administered to the other two participants. The results indicated that the test length with two thirds of the items would minimise a fatigue effect and would maximise the test reliability at the same time. The results also indicated that it would take low-proficiency learners half an hour to complete the test with two thirds of the items. Thus, the length of the entire experiment was set at 45 minutes including 15 minutes for distributing the test, explaining about the consent form, providing the instructions, and collecting the answer sheet.

Three forms were prepared for investigating the quality of all the items in the WPT using Rasch analysis. Table 45 presents the number of items for each form. Each affix was randomly included in two forms. For example, the prefix *re-* was measured in Forms A and B, while the prefix *sub-* was measured in Forms B and C. Form A shared half of the items with Form B and the rest of the items with Form C. This systematic link among the three forms was designed for linking the items in the three forms in order to estimate item difficulties in one item hierarchy and one set of person abilities including all the persons (Linacre, 2010a, p. 449).

Table 45. Number of items for each form

Section	Form A	Form B	Form C
Form	75	75	75
Meaning	48	49	49
Use	39	39	40
Total	162	163	164

The order of the three sections was determined so that the previous sections would not help answer the following sections. The form section was designed to come first, because the other two sections might help answer the items in the form section correctly. For example, if the meaning or the use of the suffix *-able* were measured earlier than its form, it would be easy to recognise the correct form from the options *-acle*, *-ague*, *-inth*, and *-able* in the form section. For the meaning and the use sections, there was no clear reason for having one section earlier than the other. Thus, for each form, the three sections were ordered in the following two ways: form-meaning-use and form-use-meaning. This design was used in order to counterbalance the order effect of the meaning and the use sections. As each of the three sections had two versions (form-meaning-use order and form-use-meaning order), a total of six forms (3 forms multiplied by 2 versions) were created. For each section, the items were randomised in order to counterbalance the order effect, but prefixes always preceded suffixes because the mixture of prefixes and suffixes might make the test confusing and affect content validity.

The test was written in a paper-based format so that the test could be administered effectively in classroom settings. For efficient data input, the answer sheet was made using an optical mark recognition (OMR) format where the participants mark their answers by darkening pre-printed circles. The information sheet, the consent form, and the instructions were translated into Japanese, the participants' L1. This ensured that even low-proficiency learners were able to fully understand the information about the

test. (See Appendix G for the six forms of the WPT used in this study.)

6.1.3 Procedure for Item Analysis

Data were collected in October and November 2010. The six test forms were randomly distributed to the participants. The data were entered into one Microsoft Office Excel 2007 (12.0.6545) spreadsheet, exported to WINSTEPS 3.71.0 (Linacre, 2010b) for Rasch analysis. In order to arrange the data in one spreadsheet, items that were not included in a form were treated as missing data. For example, the prefix *re-* was not included in Form C and thus it was treated as missing data in that form. Although this design allowed a number of missing data, researchers (Bond & Fox, 2007; Linacre, 2010a) have argued that Rasch analysis is robust with missing data which can be used intentionally by design.

Rasch analysis was performed to identify poorly written items that are misfitting to the Rasch model. As discussed in Section 4.3, items were regarded as misfit if 1) the point-measure correlation was a negative and low positive value (less than .10) or 2) the standardised fit statistics (outfit *t* and infit *t*) did not fall between -2.0 and 2.0.

A major criticism against the use of the Rasch model for analysis of the multiple-choice format is that there is no parameter accounting for lucky guessing (unexpected success by low ability respondents) (Weitzman, 1996). However, Rasch analysis can detect lucky guessing by item and person outfit statistics, and a simple strategy is to remove the lucky guesses from the data set (Wright, 1992, 1995). The subsequent section looks at whether lucky guessing was detected and how it was treated if it occurred.

6.1.4 Lucky Guessing

This section investigates the effect of lucky guessing which occurs when low ability persons unexpectedly get difficult items correct. For each section, the effect of lucky guessing was examined by item and person outfit statistics. If difficult items or low ability persons tend to be identified as misfitting, that means difficult items tend to be unexpectedly answered correctly or low ability persons tend to unexpectedly answer correctly, indicating lucky guessing. The probability of low ability persons succeeding on difficult items was also examined. If lucky guessing occurs, this success probability approaches 1/*m*, where *m* = number of multiple-choice options.

First, lucky guessing was investigated for the form section. Figure 30 illustrates the scatter plot of item difficulty and outfit *t* for this section. The horizontal axis shows item difficulty in logits, where larger numbers indicate more difficult items. The vertical axis shows outfit *t* whose values larger than 2.0 are taken as misfitting to the Rasch model. This figure indicates a tendency that difficult items are identified as misfit. Figure 31 presents the scatter plot of person ability and outfit *t*. The horizontal axis shows person ability in logits, where larger numbers indicate more able persons. The vertical axis shows outfit *t* whose values larger than 2.0 are taken as misfitting to the Rasch model.

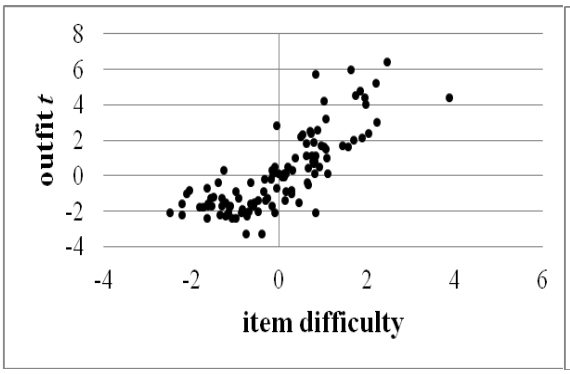


Figure 30. Item difficulty and outfit *t* for the form section

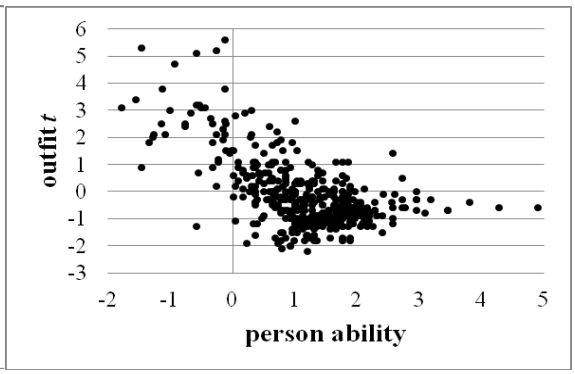


Figure 31. Person ability and outfit *t* for the form section

This figure indicates a tendency that low ability persons are identified as misfit.

Figure 32 illustrates the probability of success when a person with the ability B_n met an item with the difficulty D_i. The horizontal axis shows the difference between person ability (B_n) and item difficulty (D_i) for each response. A larger number in B_n-D_i indicates a response resulting from a person with higher ability meeting an easier item. A smaller number in B_n-D_i, on the other hand, indicates a response resulting from a person with lower ability meeting a more difficult item. The vertical axis shows the probability of a person with ability B_n succeeding on an item with difficulty D_i. The smooth line represents the theoretical model. The model predicts that the larger the B_n-D_i value is, the more likely the person is to succeed on the item, and vice versa. The dotted line, which represents the empirical data obtained from the participants in Study 1, deviates increasingly from the expected model with smaller values of B_n-D_i. In other words, when people with low ability met difficult items, their success probabilities

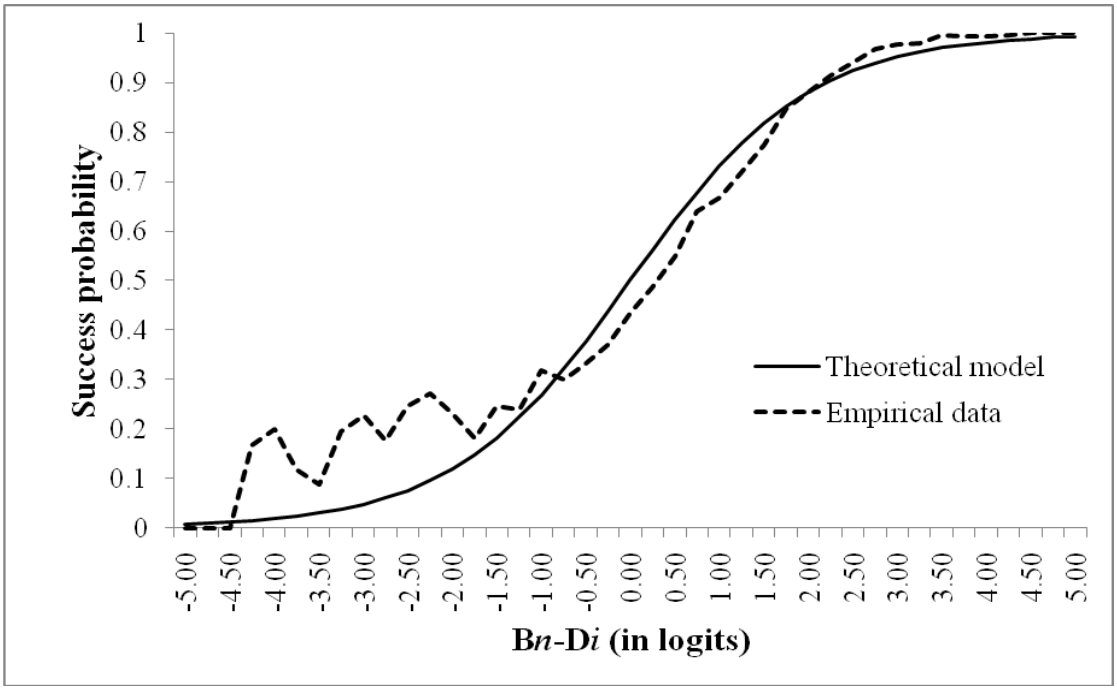


Figure 32. Success probability for the form section

approached 25% (the expected percentage of correct responses by random guessing), which was higher than the model expectation.

Taken together, Figures 30-32 indicate that lucky guessing occurred when people with low ability met difficult items in the form section. The design of the present research may allow such random guessing to occur, because a) no *Don't know* options were provided, b) the participants were asked to choose one answer even if they had no idea about the item, and c) for validation purposes all the participants worked on items with varying levels of difficulty.

Lucky guessing was corrected by deleting response records which had difficulty greater than $b + \ln(m\text{-}1)$, where b is the person's initial estimated ability and m is the number of choices (Wright & Stone, 1979). As each item had four choices, responses with an item difficulty greater than $b + 1.1$ were deleted. This presupposes "that when items are so difficult that a person can do better by guessing than by trying, then such items should not be used to estimate the person's ability" (Wright & Stone, 1979, p. 188). A total of 2,199 out of 31,255 (7.0%) responses were deleted as the result of this treatment.

Lucky guessing was also investigated for the meaning section in the same way as the form section. Figure 33 illustrates the scatter plot of item difficulty and outfit t for this section, indicating a tendency that difficult items are identified as misfit (t>2). Figure 34 presents the scatter plot of person ability and outfit t, indicating a tendency that low ability persons are identified as misfit (t>2). Figure 35 illustrates the probability of success when a person with the ability B_n met an item with the difficulty D_i, showing that the empirical data deviates increasingly from the expected model with smaller values of B_n-D_i. Taken together, Figures 33-35 indicate that lucky guessing

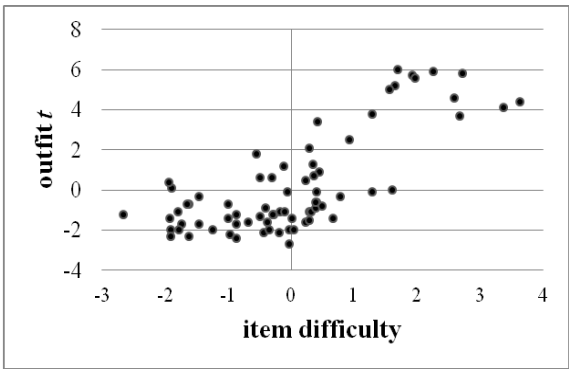


Figure 33. Item difficulty and outfit *t* for the meaning section

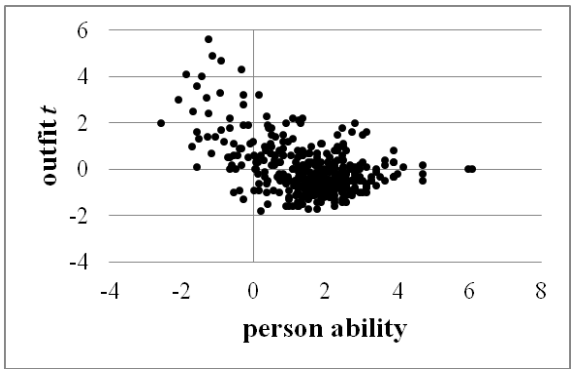


Figure 34. Person ability and outfit *t* for the meaning section

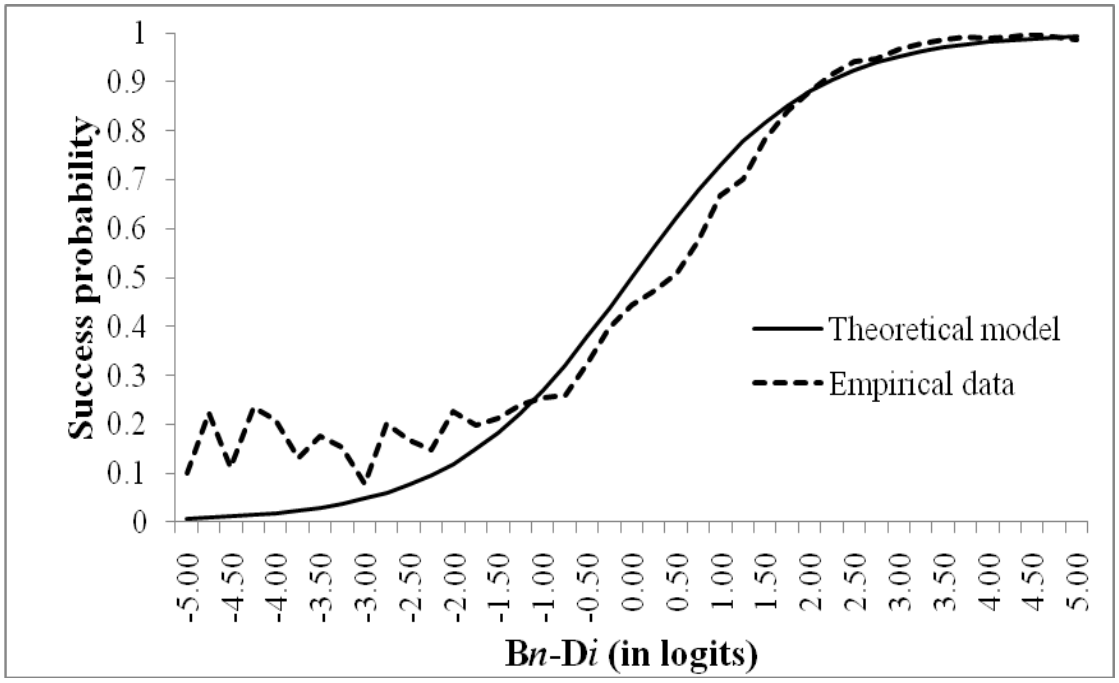


Figure 35. Success probability for the meaning section

occurred when people with low ability met difficult items in the meaning section. As with the form section, lucky guessing was corrected by deleting response records with an item difficulty greater than b + 1.1. A total of 1,597 out of 20,272 (7.9%) responses were deleted as the result of this treatment.

Finally, lucky guessing was investigated for the use section. Figure 36 illustrates the scatter plot of item difficulty and outfit *t* for this section, indicating a tendency that difficult items are identified as misfit (t>2). Figure 37 presents the scatter plot of person

ability and outfit *t*, indicating a tendency that low ability persons are identified as misfit (*t*>2). Figure 38 illustrates the probability of success when a person with the ability B_n met an item with the difficulty D_i, showing that the empirical data deviates increasingly from the expected model with smaller values of B_n-D_i.[33] Taken together, Figures 36-38

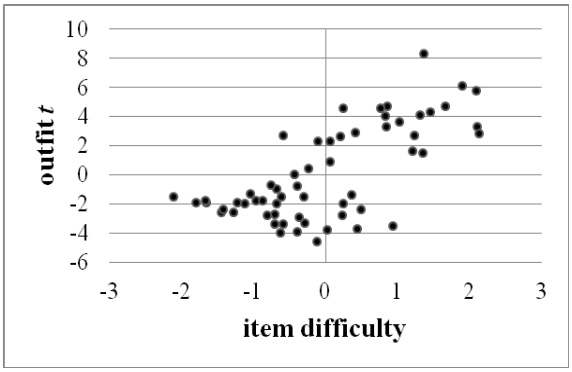


Figure 36. Item difficulty and outfit *t* for the use section

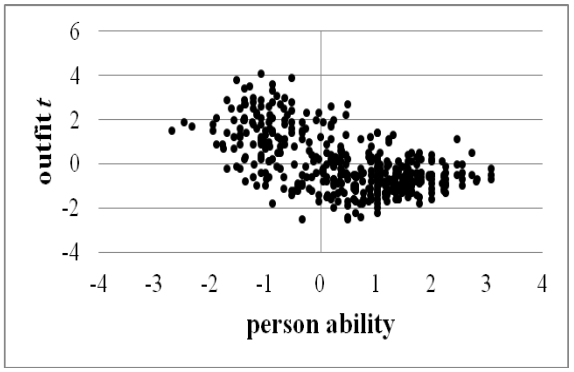


Figure 37. Person ability and outfit *t* for the use section

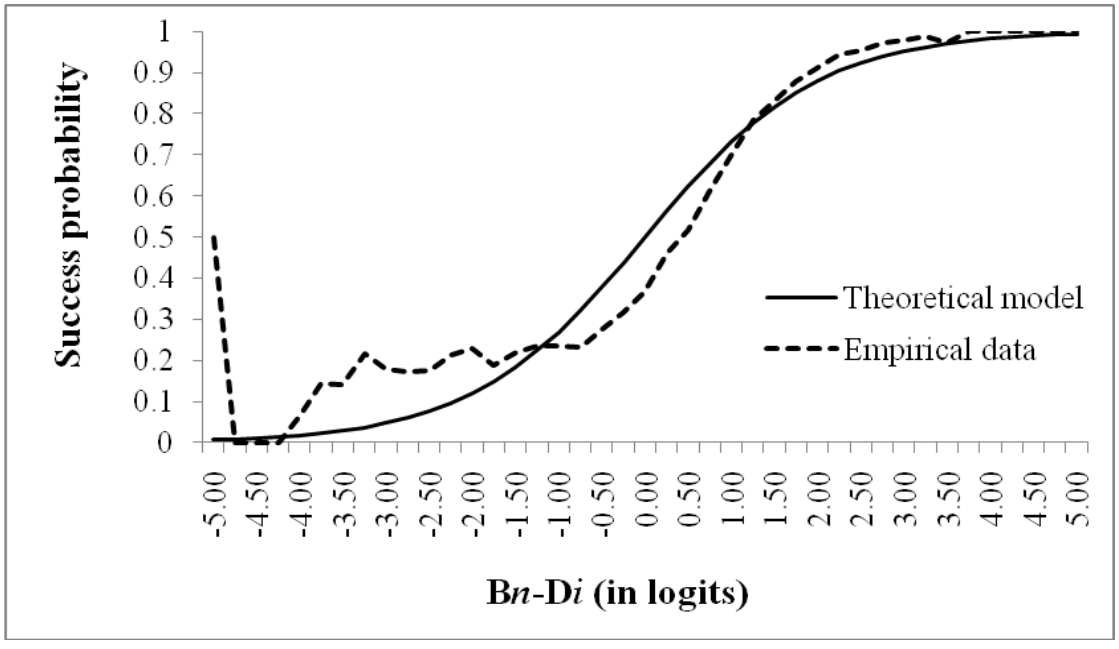


Figure 38. Success probability for the use section

[33] In Figure 38, the success probability was much higher (.5 at B_n-D_i=-5.00) than the expectation because there were only two responses (one correct and one wrong responses) at that value.

indicate that lucky guessing occurred when people with low ability met difficult items in the use section. As with the previous two sections, lucky guessing was corrected by deleting response records with an item difficulty greater than $b + 1.1$. A total of 2,721 out of 16,382 (16.6%) responses were deleted as the result of this treatment.[34]

In summary, lucky guessing was corrected for all three sections by deleting response records that have an item difficulty greater than $b + \ln(m\text{-}1)$ (Wright & Stone, 1979). The subsequent section explains the procedure for detecting and rewriting poorly written items.

6.1.5 Identifying and Rewriting Poor Items

This section aims to identify and rewrite poorly written items based on Rasch analysis. More specifically, the point-measure correlations and the Rasch fit statistics were investigated for each section. Items with negative or low positive point-measure correlations (less than .10) or items with outfit $t > 2.0$ or infit $t > 2.0$ were inspected to see if these items need to be rewritten.

Poorly written items were rewritten based on the following criteria:

1. Distractors that were chosen by many more people than the other distractors were replaced;
2. Distractors that were frequently chosen by more able people instead of the correct answer were replaced;
3. Distractors that had positive point-measure correlations were replaced; and
4. Distractors that were chosen by only a small proportion of people with low person ability estimates were replaced.

[34] A larger number of responses were deleted for the use section than the form and the meaning sections, because the use section was the most difficult and had the largest variance in person ability estimates (M=1.12, SD=1.10 for the form section, M=1.63, SD=1.49 for the meaning section, and M=0.27, SD=1.62 for the use section).

The first three of these criteria suggest that distractors may have been too close to the correct answer and prevented learners from demonstrating knowledge of affixes. A point-measure correlation is a correlation between the Rasch person ability estimates and the responses, and an option with a positive point-measure correlation indicates that more able persons tend to choose that option (Linacre, 2010a, p. 192). The correct answer should have a point-measure correlation of larger than .1, while distractors should have negative point-measure correlations. The last of the four criteria indicate that distractors may have been too easy to eliminate and not have been working well as a distractor.

6.1.5.1 Form Section

One item (Item 89: *-ling*) in the form section had a negative point-measure correlation (-.03), which indicates a need for inspecting this item. A subsequent Rasch fit analysis detected eleven items including Item 89 as underfit (outfit $t > 2.0$ or infit $t > 2.0$) and four items as overfit (outfit $t < -2.0$ or infit $t < -2.0$). Here are the details of the eleven underfit items and the procedure for rewriting them.

- Item 90: *-ling*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.99	0.13	5.6	1.38	6.2	1.29

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-ling	-wirl	-igma	-lain
% chosen	55.1	2.8	12.6	29.5
Ave. ability (logits)	1.07	-0.21	0.82	1.23
PT-measure corr.*	-.03	-.17	-.09	.14
Frequency in 10k wds**	2	1	1	10
Replaced by		**-tute**	**-reat**	**-bute**
Frequency in 10k wds		**6**	**5**	**5**

*point-measure correlation;
**frequency in the first 10,000 word families in the BNC word lists

Item 90 requires a major revision because the correct option had a negative point-measure correlation (-.03). Distractor 3 may have been most problematic because it had a positive point-measure correlation (.14) and able persons preferred this distractor (average ability = 1.23) to the correct answer (average ability = 1.07). This may be because there exists a real word *lain* (the past participle of *lie*). Another reason may be because words with the *-lain* ending are greater in number (10 word families) than words with the *-ling* ending (2 word families) in the first 10,000 word families in the BNC word lists. Some words with the *-lain* ending are highly frequent (e.g., *complain* and *explain*). Thus, this option was replaced by a less frequent word ending *-bute*. Distractor 1 was chosen by a small number of people (2.8%) with low person ability estimates (average ability = -0.21), which indicates that this option may have been too easy to eliminate. This may be because there is only one word that ends with *-wirl* (*swirl*) in the first 10,000 word families and many people may have thought that it was least likely to be a suffix. This option was replaced by *-tute* which appears more frequently than *-wirl*. Although the results indicated that Distractor 2 worked well, this option was also rewritten so that it would have a similar frequency to the other two distractors. As *stigma* (7th 1,000-word level) is the only word that ends with *-igma* in the first 10,000 word families, this option might be too easy to eliminate for future use

of the test with the rewritten options. Thus, Distractor 2 was replaced by *-reat* which is as frequent as the other distractors.

- Item 97**:** *-ous*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
-0.04	0.15	3.0	1.44	2.1	1.17

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-ous	-ail	-ope	-ime
% chosen	73.7	15.7	6.4	4.3
Ave.ability (logits)	1.25	0.97	0.48	0.93
PT-measure corr.	.16	-.07	-.13	-.07
Frequency in 10k wds	205	28	10	25
Replaced by		**-ney**		
Frequency in 10k wds		**10**		

Distractor 1 was chosen by many more people than the other two distractors. This may be because 28 words end with *-ail* in the first 10,000 word families and some of them are highly frequent (e.g., *detail* and *mail*). It could also be taken as a verb which means 'to cause problems'. This distractor was replaced by *-ney*, which is less frequent than *-ail*. Its frequency is similar to the successful distractors: Distractor 2 *-ope* has ten examples (e.g., *envelope*), and Distractor 3 *-ime* has 25 examples (e.g., *crime*) out of which 16 examples have the *-time* ending (e.g., *daytime*) in the first 10,000 word families.

- Item 93: *-most*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.96	0.13	2.9	1.22	3.7	1.17

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-most	-oard	-ogue	-laim
% chosen	55.9	15.3	12.5	16.4
Ave. ability (logits)	1.33	0.84	1.01	0.89
PT-measure corr.	.15	-.02	-.09	-.09
Frequency in 10k wds	6	20	6	5

This item may not have any problems. The three distractors obtained evenly distributed responses and the mean person ability estimate of those who chose the correct answer was higher than the mean person ability estimates of those who chose the three distractors. In addition, the correct answer showed a positive point-measure correlation, while the three distractors showed negative correlations. This item may have been identified as misfit because some of the participants suspected that *-most* was too obvious to be a correct answer. This item was not rewritten but it needs watching for future use of the test.

- Item 67: *-ess*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.58	0.14	2.6	1.21	2.8	1.15

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-ess	-ift	-ong	-nge
% chosen	61.1	10.9	14.0	14.0
Ave. ability (logits)	1.22	0.97	0.69	0.73
PT-measure corr.	.17	-.07	-.23	-.07
Frequency in 10k wds	8	12	14	25

This item may not have any problems. The three distractors obtained evenly distributed responses and the mean person ability estimate of those who chose the correct answer was higher than the mean person ability estimate of those who chose the three distractors. In addition, the correct answer showed a positive point-measure correlation,

while the three distractors showed negative correlations. This item was not rewritten but it needs watching for future use of the test.

- Item 61: *-ency*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.63	0.14	2.5	1.19	2.3	1.12

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-ency	-eeze	-eign	-yone
% chosen	59.6	13.7	20.7	6.0
Ave. ability (logits)	1.25	0.93	0.68	0.57
PT-measure corr.	.19	-.09	-.11	-.10
Frequency in 10k wds	13	5	3	2
Replaced by				**-hter**
Frequency in 10k wds				**10**

Distractor 3 was chosen by a small number of less able people. It may have been easy for more able people to eliminate because there are only two high-frequency words *anyone* and *everyone* that end with *-yone* in the 10,000 BNC word families. This distractor was replaced by *-hter*, which occurs in ten word families at varying frequency levels in the 10,000 word families (e.g., *daughter* and *laughter*).

- Item 68: *-et*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.55	0.14	2.5	1.20	2.8	1.16

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-et	-mn	-za	-ht
% chosen	62.0	22.5	2.5	13.0
Ave. ability (logits)	1.21	1.13	-0.2	0.37
PT-measure corr.	.17	.00	-.10	-.22
Frequency in 10k wds	12	6	3	74
Replaced by		**-io**	**-ob**	
Frequency in 10k wds		**12**	**15**	

This item may have been difficult, because there are 152 words that end with *-et* in the first 10,000 word families in the BNC word lists (e.g., *alphabet* and *forget*) but few of them can be divided into meaningful word parts. For this reason, some people with high ability may have thought Distractor 1 *-mn* to be more likely to be a suffix (average ability = 1.13, point-measure correlation = .00). There are six words that end with *-mn* in the first 10,000 word families (*autumn, column, condemn, damn, hymn,* and *solemn*). Some of them, for example, may have thought that *autumn* could be divided into *aut(o)* and *-mn,* and *solemn* into *sole* and *-mn*. This option was replaced by *-io* (e.g., *radio, scenario,* and *studio*) which may be less likely to be a real suffix. Distractor 2 was chosen by a small number of people with low ability. There are three words that end with *-za* in the 10,000 BNC word families (*bonanza, influenza,* and *pizza*), and from these words it may have been clear to more able people that *-za* is not a suffix. This option was replaced by *-ob* (e.g., *job, knob,* and *rob*) which is more frequent than *-za*.

- Item 50: *-ary*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.82	0.11	2.1	1.13	2.4	1.10

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-ary	-ook	-oup	-ect
% chosen	57.2	7.8	4.3	30.6
Ave. ability (logits)	1.35	0.08	0.46	1.01
PT-measure corr.	.24	-.24	-.05	-.11
Frequency in 10k wds	112	23	6	52
Replaced by				**-ech**
Frequency in 10k wds				**7**

Distractor 3 was chosen by a number of able people (average ability = 1.01), perhaps because *-ect* occurs in as many as 52 words in the 10,000 BNC word families (e.g., *collect* and *select*). The other two successful distractors have fewer examples (23 words for *-ook* and six for *-oup*). Distractor 3 *-ect* was replaced by *-ech*, which occurs in seven words in the 10,000 BNC word families (e.g., *beech* and *speech*).

- Item 62: *-ent*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.81	0.11	2.1	1.13	3.2	1.13

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-ent	-ead	-rol	-gue
% chosen	58.3	5.2	13.3	23.2
Ave. ability (logits)	1.31	0.84	0.69	0.90
PT-measure corr.	.21	-.07	-.09	-.14
Frequency in 10k wds	150	26	6	15

This item may not have any problems. Although the wrong answers were slightly biased towards Distractor 3, the three distractors were similar in the average person ability. The frequency of words with the *-gue* ending in the 10,000 BNC word families (15 words) is between the frequencies of the other two distractors. In addition, the correct answer showed a positive point-measure correlation, while the three distractors showed negative correlations. Thus, this item was not rewritten but it needs watching for future

use of the test.

- Item 56: *-dom*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.03	0.14	2.0	1.14	2.2	1.1

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-dom	-uct	-eem	-ust
% chosen	54.0	2.9	22.7	20.5
Ave. ability (logits)	1.39	1.16	0.85	0.76
PT-measure corr.	.23	.05	-.12	-.18
Frequency in 10k wds	16	11	4	24
Replaced by		**-ame**		
Frequency in 10k wds		**21**		

Distractor 1 was chosen by people with high ability (average ability = 1.16) and showed a positive point-measure correlation. Its frequency in the 10,000 BNC word families may not be problematic because the frequency of words with the *-uct* ending is between the frequencies of the other two distractors. The *-uct* ending occurs in words such as *construct*, *instruct*, and *product* which typically consist of multiple syllables so it might be mistaken for a real word part. This option was replaced by *-ame*, which is less likely to be easily mistaken for a meaningful word part, because although *-ame* occurs in more words than *-uct*, it typically attaches to monosyllabic words such as *came*, *game*, and *name*.

- Item 38: *sur-*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.79	0.14	1.8	1.14	2.6	1.14

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	sur-	att-	sco-	gue-
% chosen	59.4	25.9	8.6	6.1
Ave. ability (logits)	1.31	1.09	0.64	0.20
PT-measure corr.	.22	-.07	-.13	-.20
Frequency in 10k wds	18	11	15	4
Replaced by		**sla-**		**hal-**
Frequency in 10k wds		**15**		**11**

Distractor 3 was chosen by a small number of people with low person ability. It may have been easy to eliminate because there are only four words that start with *gue-* in the 10,000 BNC word families (*guerrilla*, *guess*, *guesswork*, and *guest*). This distractor was replaced by *hal-* (e.g., *hall* and *Halloween*), which occurs in more words than *gue-*. Distractor 1 was chosen by a large number of people with high ability. This distractor may have been popular because it was mistaken for *at-*, an allomorph of *ad-* (e.g., *attend* and *attract*). It was replaced by *sla-*, which occurs in 15 words in the 10,000 BNC word families and should not be mistaken for a real word part.

- Item 105: *-ways*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.27	0.13	1.8	1.10	2.4	1.10

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	-ways	-zard	-oice	-ypse
% chosen	50.0	27.5	16.2	6.3
Ave. ability (logits)	1.22	1.05	0.67	0.66
PT-measure corr.	.20	-.05	-.08	-.19
Frequency in 10k wds	4	4	4	1
Replaced by		**-ause**		**-ript**
Frequency in 10k wds		**6**		**6**

Distractor 3 was chosen by a relatively small number of people, perhaps because there

is only one word that ends with *-ypse* in the 10,000 BNC word families (*apocalypse*). This distractor was replaced by *-ript*, which occurs in six words in the 10,000 word families (e.g., *script* and *manuscript*). Distractor 1 was chosen by a large number of people with relatively high ability (average ability = 1.05). The four words that end with *-zard* in the 10,000 BNC word families are *haphazard*, *hazard*, *lizard*, and *wizard*. This distractor may have been mistaken for *-ard* which makes a noun such as *wizard* and *drunkard*. It was replaced by *-ause*, which has six examples in the 10,000 word families and should not be mistaken for a real word part.

The four items in Table 46 were identified as overfit based on the standardised fit statistics (outfit t < -2.0 or infit t < -2.0). Given that having less than 5% of the overfit items does not affect item and person estimates substantially (Smith Jr., 2005), it should be reasonable to conclude that these four items (3.7% of the items in the form section) do not cause serious problems; thus, these four items were not rewritten.

Table 46. Overfit items in the form section

Item No.	affix	Difficulty (logits)	S.E.	outfit *t*	outfit MNSQ	infit *t*	infit MNSQ
28	multi-	-0.45	0.16	-3.0	0.62	-2.3	0.80
27	mono-	-0.90	0.18	-2.8	0.56	-1.8	0.80
99	-some	0.89	0.13	-2.3	0.84	-2.7	0.88
7	bi-	1.78	0.14	-2.0	0.90	-2.4	0.90

6.1.5.2 Meaning Section

No item in the meaning section had a negative or a low positive point-measure correlation (less than .10). A subsequent Rasch fit analysis detected nine items as underfit (outfit t > 2.0 or infit t > 2.0) and three items as overfit (outfit t < -2.0 or infit

t < -2.0). Here are the details of the nine underfit items and the procedure for rewriting them.

- Item 65: *-ling*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.49	0.14	4.5	1.44	4.2	1.25

[Example words] earthling; underling (replaced by *weakling* and *underling*)

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	one connected with	direction	into another state/place	opposite
% chosen	55.1	10.9	24.2	9.8
Ave. ability (logits)	1.19	0.76	1.19	0.21
PT-measure corr.	.16	-.11	.00	-.13
Replaced by	**connected with**		**too much**	

Distractor 2 was chosen by a large number of people whose average person ability estimate is as high as that of those who chose the correct answer. These two options may have been too close in meaning to each other. Distractor 2 was replaced by *too much* which may be further away from the correct meaning. The correct answer *one connected with* was shortened into *connected with* in order to avoid a distinctively long option. With the example word *earthling*, Distractor 2 may be easy to eliminate because 'too much earth' does not make sense; thus, *earthling* was replaced by *weakling*.

- Item 5: *arch-*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
2.08	0.14	4.3	1.26	4.0	1.19

[Example words] archbishop; arch-rival

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	main	person/ relating to	a state of	not
% chosen	48.8	17.1	22.1	12.1
Ave. ability (logits)	1.17	1.14	1.06	1.18
PT-measure corr.	.12	-.01	-.18	.05
Replaced by		**two**	**supporting**	**earlier**

Distractor 3 was chosen by people with higher ability than the correct answer and showed a positive point-measure correlation. This distractor may have been too close in meaning to the correct answer: an archbishop is *not* a normal bishop and an arch-rival is *not* a normal rival. Distractor 4 was replaced by *earlier* which is further away from the correct meaning. Distractors 1 and 2 were also chosen by people with relatively high ability. They may have thought that an option carrying the broadest meaning would be most likely to be the correct answer. Distractors 1 and 2 were replaced by *two* and *supporting*, which are less vague than *person/relating to* and *a state of*.

- Item 54: *-fold*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.55	0.16	3.6	1.59	3.4	1.29

[Example words] two-fold; three-fold

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	times	under	not	self
% chosen	70.2	16.5	3.5	9.8
Ave. ability (logits)	1.16	0.83	-0.04	1.04
PT-measure corr.	.20	-.10	-.14	-.11
Replaced by			**over**	

Distractor 2 was chosen by only a small number of people with low ability. As the

example words refer to the notion of number, it may have been easy to eliminate this option. This distractor was replaced by *over*, which would be consistent with the example words (*over two* and *over three*).

- Item 31: *pro-*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.95	0.14	2.9	1.18	3.2	1.15

[Example words] pro-democracy; pro-life

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	supporting	without	a state of	too much
% chosen	50.0	8.6	22.9	18.6
Ave. ability (logits)	1.20	1.28	1.07	1.01
PT-measure corr.	.18	-.08	-.09	-.09
Replaced by		**against**	**one**	

Distractor 1 was chosen by a relatively small number of people, but their average person ability estimate was higher than that of those who chose the correct answer. They may have recalled another meaning such as 'substitute for' (e.g., *procathedral* and *pronoun*). This distractor was replaced by *against*, which is the opposite of the correct answer. Distractor 2 was chosen by a large number of people with relatively high ability. This may have been due to their use of the test-taking strategy to choose the broadest meaning. This distractor was replaced by *one*, which is less vague than *a state of*.

- Item 52: *-et*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
2.22	0.15	2.6	1.18	2.7	1.13

[Example words] packet; owlet

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	small	new	can be	not
% chosen	43.0	7.4	46.8	2.8
Ave. ability (logits)	1.21	0.73	1.00	0.21
PT-measure corr.	.21	-.08	-.18	-.03
Replaced by			**supporting**	

Distractor 2 was chosen by a large number of people. This may be because *-et* was mistaken for *-ed* (e.g., *packed* rather than *packet*) and *can be* was taken as a passive voice. This distractor was replaced by *supporting*, which should not be mistaken for another word part.

- Item 33: *semi-*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
-0.6	0.20	2.4	1.74	3.1	1.46

[Example words] semi-final; semi-skilled

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	half	person/thing	direction	can be
% chosen	83.8	2.9	8.6	4.7
Ave. ability (logits)	1.20	0.40	0.58	1.37
PT-measure corr.	.27	-.28	-.17	-.02
Replaced by				**beyond**

Distractor 3 was chosen by people with higher person ability estimates than those who chose the correct answer. This may be because they thought that a semi-final is very close to a final and *can be* taken as a final. This distractor was replaced by *beyond*, which might be further away from the correct meaning than *can be*.

- Item 13: *ex-*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
3.2	0.17	2.3	1.18	2.4	1.16

[Example words] ex-wife; ex-member

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	earlier	person	without	can be
% chosen	30.2	10.8	47.8	11.2
Ave. ability (logits)	1.26	0.46	1.22	1.04
PT-measure corr.	.15	-.10	-.05	-.10
Replaced by			**bad**	

Distractor 2 was chosen by a large number of people with high ability. This may be because they associated an ex-wife with a man *without* his wife. This distractor was replaced by *bad*, which is further away from the correct meaning than *without*.

- Item 37: *trans-*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.41	0.17	2.1	1.37	1.9	1.18

[Example words] transform; transplant

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	into another state/place	two	one connected with	a state of
% chosen	73.7	4.0	19.1	3.2
Ave. ability (logits)	1.26	0.37	0.97	-0.01
PT-measure corr.	.31	-.23	-.12	-.31
Replaced by			**main**	**too much**

Distractor 2 was chosen by a number of people with relatively high ability, perhaps because the meaning was too close to the correct answer. This distractor was replaced by *main*, which is further away from the correct meaning than *one connected with*.

Distractor 3 was chosen by a small number of people with low person ability. They may have found it difficult to differentiate between *a state of* and *into another state/person*, as both refer to the notion of *state*. This distractor was replaced by *too much*, which is further away from the correct meaning than *a state of*.

- Item 73: *-wise*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.95	0.14	1.4	1.10	2.1	1.10

[Example words] clockwise; stepwise

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	direction	person	against	one
% chosen	52.0	13.0	21.7	13.3
Ave. ability (logits)	1.28	0.94	0.83	1.21
PT-measure corr.	.27	-.01	-.21	-.17
Replaced by			**new**	

Distractor 2 was chosen by a number of people, perhaps because *against* could be taken as a kind of *direction*. This option was replaced by *new*, which would be further away from the correct meaning than *against*.

The three items in Table 47 were identified as overfit based on the standardised fit statistics (outfit $t < -2.0$ or infit $t < -2.0$). Given that having less than 5% of the overfit items does not affect item and person estimates substantially (Smith Jr., 2005), it should be reasonable to conclude that these four items (4.1% of the items in the meaning section) do not cause serious problems; thus, these three items were not rewritten.

Table 47. Overfit items in the meaning section

Item No.	affix	Difficulty (logits)	S.E.	outfit *t*	outfit MNSQ	infit *t*	infit MNSQ
26	multi-	-0.01	0.17	-2.6	0.57	-2.6	0.76
45	-ee	-0.93	0.21	-2.2	0.45	-1.1	0.85
7	bi-	0.49	0.16	-2.1	0.72	-2.2	0.83

6.1.5.3 Use Section

One item (Item 27: *-ent* (noun)) in the use section had a negative point-measure correlation (-.27), which indicates a need for inspecting this item. A subsequent Rasch fit analysis detected eleven items including Item 27 as underfit (outfit t > 2.0 or infit t > 2.0) and ten items as overfit (outfit t < -2.0 or infit t < -2.0). Unlike the previous two sections, all the items have the same options in the use section (*noun, verb, adjective,* and *adverb*). Thus, what could be done is to examine whether example words were appropriate or not. Here are the details of the eleven underfit items and the procedure for rewriting them.

- Item 28: *-ent* (noun)

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.94	0.22	6.6	1.90	6.2	1.53

[Example words] referent; respondent (replaced by *president* and *respondent*)

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	noun	verb	adjective	adverb
% chosen	38.2	9.0	42.4	10.4
Ave. ability (logits)	1.17	0.76	1.21	0.33
PT-measure corr.	-.23	-.13	.27	-.18

The correct answer was *noun*, but many more people with higher ability estimates chose

adjective instead of the correct answer. The suffix *-ent* frequently forms an adjective (e.g., *different*), and in most cases nouns with the *-ent* ending can also be adjectives (e.g., *resident*). In the BNC, while all the 1,602 examples of *respondent* are nouns (no adjective examples), 19 out of 268 (7.1%) examples of *referent* are used as adjectives. *Referent* was selected as an example word because it belonged to the most frequent word level (1st 1,000); however, it would be better to replace it with *president*, which has no adjective examples and occurs more than 10,000 times in the BNC although it belongs to the 4th 1,000-word level.

- Item 58: *-y* (adjective)

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.55	0.15	4.5	1.50	5.4	1.38

[Example words] lucky; healthy

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	51.6	38.2	3.9	6.3
Ave. ability (logits)	1.17	1.08	-0.33	0.76
PT-measure corr.	.18	.05	-.15	-.18

The correct answer was *adjective*, but a large number of people chose *noun*. The positive point-measure correlation indicates a problem with this distractor. This may be because the suffix *-y* can also make a noun (e.g., *difficulty*). However, the example words can only be taken as adjectives: no noun examples of *lucky* and *healthy* are found in the BNC. This item was not changed but needs watching for future use.

- Item 2: *be-*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.30	0.16	4.1	1.35	4.5	1.27

[Example words] <u>be</u>little; <u>be</u>friend

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	verb	noun	adjective	adverb
% chosen	40.0	14.0	35.8	10.2
Ave. ability (logits)	1.34	0.92	0.98	0.34
PT-measure corr.	.10	-.06	-.03	-.10

The correct answer was *verb*, but many people chose *adjective*. This may be because some words are typically used in past participle form (e.g., *beloved* and *bemused*). However, the example words should be appropriate because they cannot be taken as adjectives. This item was not changed but needs watching for future use.

- Item 12: *-ant* (noun)

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
-0.73	0.16	3.8	1.68	4.4	1.39

[Example words] consult<u>ant</u>; serv<u>ant</u>

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	noun	verb	adjective	adverb
% chosen	68.2	7.5	18.9	5.4
Ave. ability (logits)	1.26	0.59	1.17	0.56
PT-measure corr.	.26	-.23	-.05	-.20

The correct answer was *noun*, but a large number of people with high ability chose *adjective*. This may be because the suffix *-ant* can also make an adjective (e.g., *expectant*), but the example words can only be taken as nouns: no adjective examples of

consultant and *servant* are found in the BNC. This item was not changed but needs watching for future use.

- Item 23: *-en* (adjective)

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.05	0.16	3.6	1.48	4.1	1.35

[Example words] wooden; golden

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	57.9	12.6	21.6	7.9
Ave. ability (logits)	1.39	0.79	1.01	0.10
PT-measure corr.	.30	-.31	.07	-.34

The answer was *adjective*, but a large number of people with high ability chose *verb*. The positive point-measure correlation indicates a problem with this distractor. This may be because the suffix *-en* can also make a verb (e.g., *darken*), but the example words can only be taken as adjectives: no verb examples of *wooden* and *golden* are found in the BNC. This item was not changed but needs watching for future use.

- Item 49: *-ory*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.21	0.16	3.3	1.29	3.1	1.18

[Example words] sensory; contradictory

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	41.8	33.3	5.6	19.3
Ave. ability (logits)	1.19	1.08	0.19	0.92
PT-measure corr.	.22	-.06	-.14	-.18

The correct answer was *adjective*, but a large number of people with high ability chose *noun*. Some of them may have confused *-ory* with *-ry* which indicates a noun (e.g., *jewelry* and *rivalry*). Some words with the *-ory* ending can be used as a noun as well as an adjective (e.g., *auditory*). However, the example words can only be taken as adjectives: no noun examples of *sensory* and *compensatory* are found in the BNC. This item was not changed but needs watching for future use.

- Item 6: *-age*

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
-0.12	0.15	3.1	1.39	4.0	1.31

[Example words] short<u>age</u>; cover<u>age</u>

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	noun	verb	adjective	adverb
% chosen	59.4	14.6	18.9	7.1
Ave. ability (logits)	1.33	0.70	1.10	0.65
PT-measure corr.	.29	-.17	-.07	-.24

The correct answer was *noun*, but people with relatively high ability chose *adjective*. This item may not be problematic, because wrong answers were not extremely biased towards one distractor, people who chose the correct answer had the highest average person ability estimate, and all the distractors had negative point-measure correlations. Thus, this item was not changed but needs watching for future use.

- Item 14: *-ary* (adjective)

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.36	0.15	3.1	1.35	3.7	1.26

[Example words] revolution<u>ary</u>; parliament<u>ary</u>

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	52.3	16.5	4.6	26.7
Ave. ability (logits)	1.23	1.11	0.26	0.79
PT-measure corr.	.27	.02	-.12	-.29

The correct answer was *adjective*, but people with relatively high ability chose *noun*. The positive point-measure correlation indicates a problem with this distractor. This may be because some words with the *-ary* ending are nouns (e.g., *secretary*). However, the example words can only be taken as adjectives: no noun examples of *revolutionary* and *parliamentary* are found in the BNC. Thus, this item was not changed but needs watching for future use.

- Item 17: *-ate* (verb)

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
0.57	0.16	2.9	1.29	3.2	1.23

[Example words] formul<u>ate</u>; activ<u>ate</u>

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	verb	noun	adjective	adverb
% chosen	51.4	12.9	24.8	10.8
Ave. ability (logits)	1.43	0.71	0.95	0.63
PT-measure corr.	.27	-.16	-.06	-.26

The correct answer was *verb*, but a number of people chose *adjective*. This may be because some words with the *-ate* ending are adjectives (e.g., *passionate*). However, the example words can only be taken as adjectives: no noun examples of *formulate* and *activate* are found in the BNC. This item was not changed but needs watching for future

use.

- Item 44: *-ly* (adjective)

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
2.62	0.19	2.8	1.43	0.8	1.08

[Example words] manly; friendly (replaced by *lively* and *friendly*)

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	31.2	4.2	3.2	61.4
Ave. ability (logits)	0.94	0.19	0.64	1.18
PT-measure corr.	.18	-.12	-.16	.00

The correct answer was *adjective*, but quite a few people with high ability chose *adverb*. This may be because the majority of words with the *-ly* ending are adverbs (e.g., *widely*). For an example word, *manly* was chosen because it is listed in the first 1,000 word families in the BNC word lists, but *manly* itself (including its inflective forms *manly*, *manlier*, and *manliest*) is not very frequent (123 occurrences in the BNC). It might have been mistaken for much more frequent adverbs such as *mainly*. This example word was replaced by *lively*, which is also listed in the first 1,000 word families and occurs 1,529 times (including its inflective forms *lively*, *livelier*, and *liveliest*) in the BNC.

- Item 16: *-ate* (adjective)

[Statistics]

Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
1.83	0.17	2.1	1.18	1.9	1.12

[Example words] passionate; determinate (replaced by *passionate* and *fortunate*)

[Options]

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	33.0	12.9	43.0	11.1
Ave. ability (logits)	1.24	1.00	1.22	0.77
PT-measure corr.	.19	-.17	-.01	-.20

The correct answer was *adjective*, but a large number of people with high ability chose *verb*. This may be because some words with the *-ate* ending are verbs (e.g., *activate*). Another reason may be because one of the example words *determinate* was not familiar with the test-takers (118 occurrences in the BNC). This example word was replaced by *fortunate*, which is more frequent (1,263 occurrences) than *determinate*.

The ten items in Table 48 were identified as overfit based on the standardised fit statistics (outfit t < -2.0 or infit t < -2.0) which may potentially identify a number of good items as misfit with a large sample size (Karabatsos, 2000; Linacre, 2003; Smith, et al., 2008). However, the unstandardised statistics indicated that only two items (Items 34 and 24) had the outfit mean-square statistics smaller than 0.70 which may be taken as unacceptable values (Bond & Fox, 2007; Linacre, 2002). No infit mean-square

Table 48. Overfit items in the use section

Item No.	Affix	Difficulty (logits)	S.E.	Outfit *t*	Outfit MNSQ	Infit *t*	Infit MNSQ
4	en-	0.26	0.15	-3.3	0.71	-3.8	0.77
34	-ify	-0.31	0.16	-3.2	0.62	-3.2	0.75
59	-y (n)	0.37	0.16	-3.2	0.70	-3.4	0.78
3	em-	0.53	0.16	-3.1	0.72	-3.3	0.81
10	-ancy	-0.10	0.16	-2.6	0.70	-2.5	0.81
53	-ty	0.87	0.16	-2.4	0.82	-2.6	0.85
24	-ence	-0.93	0.16	-2.3	0.65	-3.2	0.75
46	-ness	-0.47	0.15	-2.2	0.74	-2.7	0.80
5	-able	-0.85	0.17	-2.1	0.70	-3.0	0.75
39	-ity	-0.46	0.16	-2.1	0.71	-2.0	0.83

statistics are smaller than 0.70. Given that standardised fit statistics are highly susceptible to sample size and having less than 5% of the overfitting items does not affect item and person estimates substantially (Smith Jr., 2005), it should be reasonable to conclude that these two items (3.4% of the items in the use section) do not cause serious problems; thus, the overfit items in Table 48 were not changed.

In summary, Rasch analysis detected eleven misfit items for the form section, nine for the meaning section, and eleven for the use section. Table 49 summarises the misfit items identified. The WPT was revised by rewriting these items. The subsequent section explains the study that investigated the quality of the revised WPT.

Table 49. Summary of misfit items in the WPT

Affix	Form	Meaning	Use	Affix	Form	Meaning	Use
arch-		✓		-ency	✓		
be-			✓	-ent (n)	✓		✓
ex-		✓		-ess	✓		
pro-		✓		-et	✓	✓	
semi-		✓		-fold		✓	
sur-	✓			-ling	✓	✓	
trans-		✓		-ly (a)			✓
-age			✓	-most	✓		
-ant (n)			✓	-ory			✓
-ary	✓		✓	-ous	✓		
-ate (a)			✓	-ways	✓		
-ate (v)			✓	-wise		✓	
-dom	✓			-y (a)			✓
-en (a)			✓				

6.2 Study 2

The purpose of Study 2 was to empirically examine the quality of the revised WPT based on data from participants with various L1 backgrounds.

6.2.1 Participants

The revised WPT was written in a web-based format in order to effectively collect data from people with various L1 backgrounds all over the world. A total of 1,348 people (470 males, 580 females, and 298 unspecified) participated in the research.[35] Their ages ranged between 10 and 73, with the average being 29.4 (SD=11.9). The participants were recruited in the following way: 22 participants took the test under the supervision of a teacher as part of their English classroom activities, 76 participants took it because their English teachers recommended the test to their students, and the other participants knew the test through online social networking services where people recommended the test to their friends (e.g., Facebook) and online advertisements where the advertisement of the test was displayed on a web page along with search results when someone searched using one of the pre-determined keywords (Google AdWords). The test was taken by participants with a wide variety of L1s (Table 50) from more than 100 countries (Table 51). This may indicate that any advantages or disadvantages from cognates and loan words for one native language over another are counterbalanced. The participants also varied widely in their proficiency levels. Their vocabulary size as measured by Nation and Beglar's (2007) Vocabulary Size Test from 62 of the participants ranged between 3,200 and 13,100 word families (Mean=8,958, SD=2,326). This may indicate that the results from these participants are highly generalisable to different groups of people.

[35] Although a total of 1,439 people completed the test, the data from 1,348 people were used for analysis. Data from 91 people were excluded from the analysis because the response record showed that they completed the test too quickly (less than 1 second per item) or too slowly (more than 15 seconds per item). Test-takers who gave too quick answers typically got about 25% of the items correct, indicating that they did the test without thinking carefully. Four people spent more than 15 seconds per item and their responses showed an irregular pattern in terms of Rasch fit statistics, perhaps because they relied on external resources such as a dictionary for some difficult items.

Table 50. Participants' L1s

L1	No. of participants	%	L1	No. of participants	%
English	226	16.8	Malayalam	8	0.6
Arabic	102	7.6	Italian	7	0.5
Hindi	93	6.9	Ukrainian	7	0.5
Japanese	86	6.4	Telugu	6	0.4
Urdu	67	5.0	Malay	5	0.4
Russian	61	4.5	Marathi	5	0.4
Indonesian	60	4.5	Panjabi, Punjabi	5	0.4
Filipino	53	3.9	Burmese	4	0.3
Spanish	47	3.5	Korean	4	0.3
Vietnamese	47	3.5	Rajasthani	4	0.3
Chinese	46	3.4	Gujarati	3	0.2
Tamil	34	2.5	Finnish	2	0.1
Tagalog	28	2.1	Javanese	2	0.1
German	21	1.6	Kannada	2	0.1
French	19	1.4	Pashto, Pushto	2	0.1
Persian	14	1.0	Sindhi	2	0.1
Polish	14	1.0	Yoruba	2	0.1
Portuguese	14	1.0	Awadhi	1	0.1
Turkish	14	1.0	Bhojpuri	1	0.1
Bengali	11	0.8	Czech	1	0.1
Thai	11	0.8	Estonian	1	0.1
Romanian, Moldavian, Moldovan	10	0.7	Uzbek	1	0.1
Dutch	8	0.6	Unspecified/other	187	13.9

6.2.2 Materials

The test was written in a web-based format so that the test could be taken effectively by people all over the world. The web-based format has a number of advantages over a paper-based one. First, the participants can take the test anywhere, anytime when they have access to the Internet. Second, the response time can be recorded for each item so that responses without careful thinking (too short response time) and responses using external resources such as a dictionary (too long response time) may be excluded from the analysis. Third, the web-based format makes sure that the test-takers cannot go back

Table 51. Locations of the participants (five or more participants)

Location	No. of participants	%	Location	No. of participants	%
India	148	11.0	Algeria	9	0.7
United States	120	8.9	Ethiopia	9	0.7
Philippines	92	6.8	France	9	0.7
New Zealand	72	5.3	Mexico	9	0.7
Japan	69	5.1	Nepal	9	0.7
Pakistan	69	5.1	Netherlands	9	0.7
Indonesia	57	4.2	Serbia	9	0.7
Egypt	46	3.4	Bulgaria	8	0.6
United Kingdom	36	2.7	Hong Kong	8	0.6
Vietnam	35	2.6	Kazakhstan	8	0.6
Russian Federation	30	2.2	United Arab Emirates	8	0.6
Canada	29	2.2	Argentina	7	0.5
Bangladesh	18	1.3	Australia	7	0.5
China	18	1.3	Iran	7	0.5
Poland	16	1.2	Lebanon	7	0.5
Thailand	16	1.2	Spain	7	0.5
Cambodia	13	1.0	Ukraine	7	0.5
Mongolia	13	1.0	Hungary	6	0.4
Georgia	12	0.9	Morocco	6	0.4
Iraq	12	0.9	Albania	5	0.4
Malaysia	12	0.9	Armenia	5	0.4
Singapore	12	0.9	Belgium	5	0.4
Turkey	12	0.9	Bosnia and Herzegovina	5	0.4
Germany	11	0.8	Italy	5	0.4
Saudi Arabia	11	0.8	Lithuania	5	0.4
Sri Lanka	11	0.8	Norway	5	0.4
Brazil	10	0.7	Romania	5	0.4
Honduras	10	0.7	Trinidad and Tobago	5	0.4
Taiwan	10	0.7	Other	134	9.9

to the previous questions nor skip any questions. Finally, an order effect is completely counterbalanced because the order of the items and the options is automatically randomised for each test-taker.

The test length was determined based on Study 1 so that the test would achieve an estimated Rasch person reliability of .9 which indicates that the test discriminates the sample into three or four levels (Linacre, 2010a, p. 512). The number of items required to arrive at a reliability of .9 was estimated by the following Spearman-Brown prediction formula (Brown, 1910; Spearman, 1910):

$$T = C \times \frac{R_T\,(1 - R_C)}{R_C(1 - R_T)},$$

where T = target number of items, C = current number of items, R_T = target person reliability, and R_C = current person reliability. For each form used in Study 1, the number of items for arriving at the reliability of .9 was estimated after the deletion of misfit items (Table 52).[36]

Table 52. Estimated number of items (reliability = .9)

	Form section			Meaning section			Use section		
Form	A	B	C	A	B	C	A	B	C
Current No. of items	66	69	66	43	44	42	30	34	33
Reliability	.92	.91	.92	.88	.87	.89	.90	.90	.92
Target No. of items	51.7	61.4	51.7	52.8	59.2	46.7	30.0	34.0	25.8
Ave. target No. of items	54.9			52.9			29.9		

For the form section, the average target number of items was 54.9, indicating that at least 55 items would be needed to arrive at the reliability of .9. In the same way, 53

[36] Persons with extreme scores were included in the analysis (two persons got all items correct for the meaning section) and model reliability instead of real reliability was used based on Linacre's (2010a) following suggestion: "in general, Cronbach Alpha overestimates reliability, Rasch underestimates it. So, when it is likely that the Rasch reliability will be compared with conventional KR-20 or Cronbach Alpha reliabilities, […] then include extreme persons and report the higher Rasch reliability, the "Model" reliability, computed on the assumption that all unexpectedness in the data is in accord with Rasch model predictions" (p.512).

items would be needed for the meaning section, and 30 items for the use section.[37]

The test was designed using a common item linking method where all test forms shared particular items in common in order to make sure that each form contains items with good fit statistics that would be useful for linking items in different forms (Bond & Fox, 2007; Linacre, 2010a; Wright & Stone, 1979). For each section, ten or more items with good fit statistics were chosen as common items because a common item linking method requires at least five items that are spread out across the difficulty continuum (Linacre, 2010a, p. 450). Here are the criteria for choosing the common items. The information on the items (e.g., difficulty estimates and fit statistics) is based on the results of Study 1.

1. Common items were selected from items with a wide range of difficulty estimates.
2. Common items were selected from items that showed good fit indices (outfit mean-squares ranging between 0.8 and 1.2).
3. Common items were selected from items that showed invariance in difficulty estimates between high-ability and low-ability groups.
4. Difficulty estimates for common items should not be affected by knowledge of loan words in Japanese.

In order to meet the first criterion, stratified sampling was conducted; that is, items were classified into difficulty levels each with a one-logit range (e.g., between 1 and 2 logits) and were chosen from each level. Second, the outfit mean-square range of 0.8-1.2 may be reasonable for common items because items within this range may be useful for multiple-choice tests of high stakes (Bond & Fox, 2007, p. 243). Third, ideal common items should be invariant in difficulty estimates across samples (Bond & Fox, 2007). The sample was divided into equally-sized sub-samples with one being a high-ability

[37] Missing data may decrease reliability (Linacre, 2010a), but very few people left items unanswered and omitting people with missing data did not improve reliability.

group and the other being a low-ability group. The common items showed similar item difficulty estimates (non-significant DIF) between the high- and low-ability groups. The last criterion was set up because different samples with different L1 backgrounds might show different item difficulty estimates for items that may be affected by knowledge of loan words in Japanese. For example, in Japanese *anchi-* (*anti-*) is often used to create words with the meaning of *against* such as *anchi-kyojin* (*anti-Giants*), which might lead to the underestimation of the item difficulty of *anti-* with Japanese learners. In fact, Mochizuki and Aizawa (2000) indicated that Japanese learners may be able to demonstrate their knowledge of some infrequent affixes based on their knowledge of frequent loan words in Japanese. In addition to the items with good fit statistics, each form also included the misfit items identified in Study 1 in order to obtain as much information as possible for examining the quality of the rewritten items. In sum, each form was designed to include 1) items with good fit statistics for common item linking and 2) the misfit items identified in Study 1.

For each test-taker, the test was programmed to have all the common items and to randomly select items from the other items. Table 53 presents the number of items included in each test form. The number of items was determined based on the following criteria:

1. In order to achieve an estimated reliability of .9, the form section had at least 55 items, the meaning section had at least 53 items, and the use section had at least 30 items;
2. Ten or more items with good fit statistics were included for common item linking;
3. All misfit items were included in every form; and
4. The rest of the items (other than items for common item linking and misfit items) had a 50% chance of being selected so that each item would be answered by half of the participants. Given that at least 250 examinees are needed for stable item calibrations with 99% confidence (Linacre, 1994), this design

required at least 500 participants, which was considered to be highly achievable.

Table 53. Number of items for each form of the revised WPT

Section	Common items		Other items	Total
	Misfit items	Good items		
Form	11	10	44	65
Meaning	9	24	21	54
Use	11	10	19	40

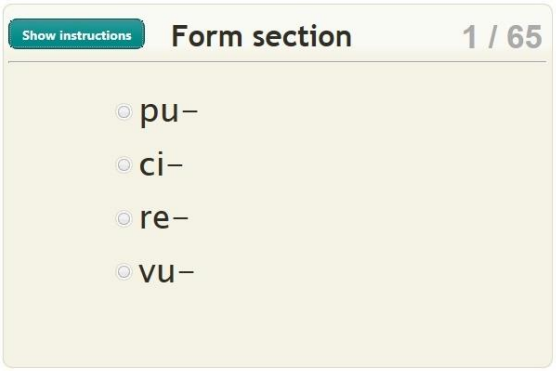
Table 53 shows that each form has a total of 159 items, which indicates that the web-based WPT was slightly shorter than the paper-based WPT used in Study 1 and thus the test was expected to be completed within 30 minutes.

The order of the three sections (form, meaning, and use) was determined so that the previous sections would not help answer the following sections. The form section always came first because the other two sections might help answer the items in the form section correctly. The order of the other two sections (meaning and use) was randomised for each test-taker because there was no clear reason for having one section earlier than the other. Thus, for each form, the three sections were randomly ordered in the following two ways: form-meaning-use and form-use-meaning. For each test-taker, the item order was randomised in order to counterbalance an order effect, but prefixes were always followed by suffixes because the mixture of prefixes and suffixes might make the test confusing. For the form and the meaning sections, the order of the options was also randomised for each test-taker in order to counterbalance an order effect. The option order in the use section was not randomised because all the items in this section had the fixed four options (noun, verb, adjective, and adverb) and randomised options might increase construct-irrelevant difficulty.

Test-takers were presented with one item on a computer screen at a time, and when they clicked on an answer they were presented with the next item. They were not allowed to go back to the previous items to change the answers in order to make sure that the responses were not affected by the subsequent items. They could read the instructions at any time if they clicked on the *Show Instructions* button which was presented at the upper left of every item. They were also presented with an indication of their progress: the number of items they have completed and the total number of items in the section. Figures 39-41 illustrate the examples of the web-based WPT for the three sections. Each figure presents two examples: one for a prefix and the other for a suffix. For the form section, test-takers must choose a real affix from four options (Figure 39). For the meaning section, they must choose the closest meaning of the target affix (Figure 40). For the use section, they must choose the part of speech that the target affix forms (Figure 41). The response time was recorded for each item so that unreliable data could be identified.

For the participants who completed the test, a report on their level of word part knowledge was provided as soon as they finished the test. They were also presented with a list of word parts that would help them to improve their knowledge of word parts.

Example 1



Example 2

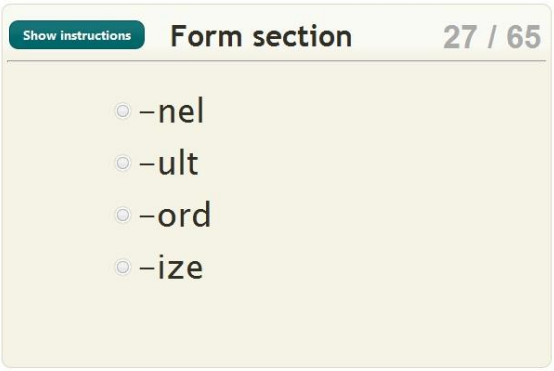
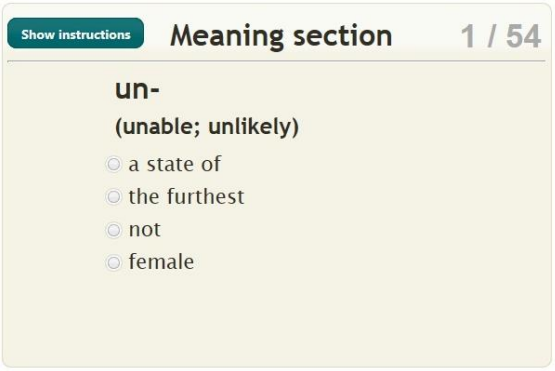


Figure 39. Examples of the web-based form section

Example 1



Example 2

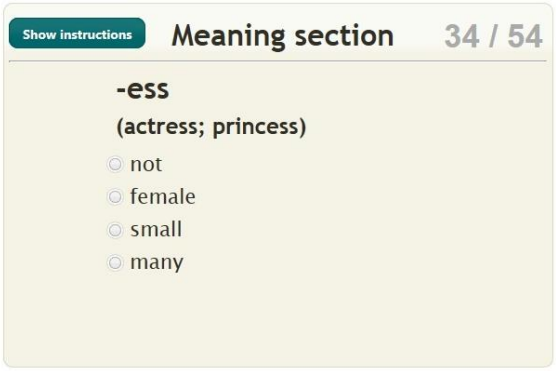
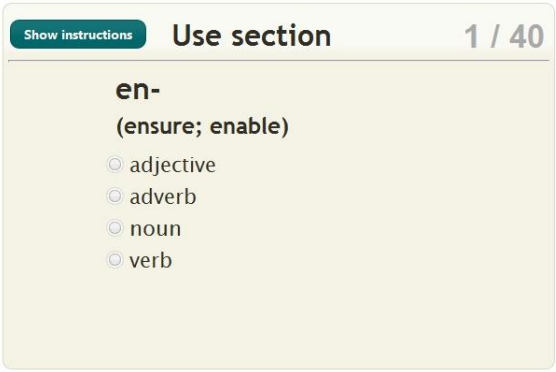


Figure 40. Examples of the web-based meaning section

Example 1



Example 2

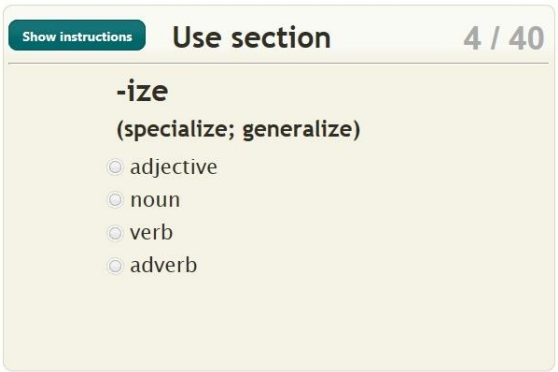


Figure 41. Examples of the web-based use section

6.2.3 Procedure for Item Analysis

Data were collected through the Internet between July and October 2011. The data were entered into one Microsoft Office Excel 2007 (12.0.6545) spreadsheet, exported to WINSTEPS 3.71.0 (Linacre, 2010b) for Rasch analysis. As with Study 1, items that were not taken by a test-taker were treated as missing data.

As with Study 1, the effect of lucky guessing was investigated in order to examine whether lucky guessing should be corrected for item analysis. In so doing, Rasch item and person outfit statistics were examined for each section. If difficult items or low ability persons tend to be identified as misfit, that means difficult items tend to be unexpectedly answered correctly or low ability persons tend to unexpectedly answer correctly, indicating lucky guessing. The probability of low ability persons succeeding on difficult items was also examined. If lucky guessing occurs, this success probability approaches 1/*m*, where *m* = number of multiple-choice options.

First, the effect of lucky guessing was investigated for the form section. Figure 42 illustrates the scatter plot of item difficulty and outfit *t* for this section. The horizontal axis shows item difficulty in logits, where larger numbers indicate more difficult items. The vertical axis shows outfit *t* whose values larger than 2.0 are taken as misfitting to the Rasch model. This figure indicates a tendency that difficult items are identified as misfit. Figure 43 presents the scatter plot of person ability and outfit *t*. The horizontal axis shows person ability in logits, where larger numbers indicate more able persons. The vertical axis shows outfit *t* whose values larger than 2.0 are taken as misfitting to the Rasch model. This figure indicates a tendency that low ability persons are identified as misfit.

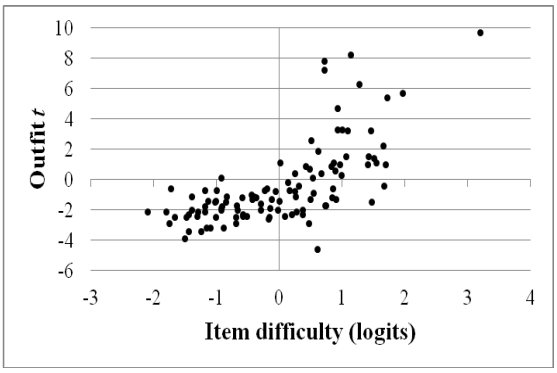


Figure 42. Item difficulty and outfit *t* for the form section

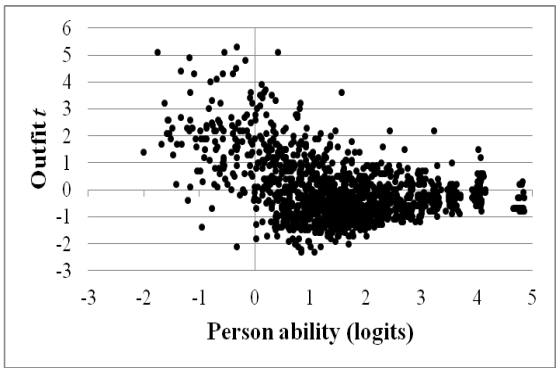


Figure 43. Person ability and outfit *t* for the form section

Figure 44 illustrates the probability of success when a person with the ability B_n met an item with the difficulty D_i. The horizontal axis shows the difference between person ability (B_n) and item difficulty (D_i) for each response. A larger number in B_n-D_i indicates a response resulting from a person with higher ability meeting an easier item. A smaller number in B_n-D_i, on the other hand, indicates a response resulting from a

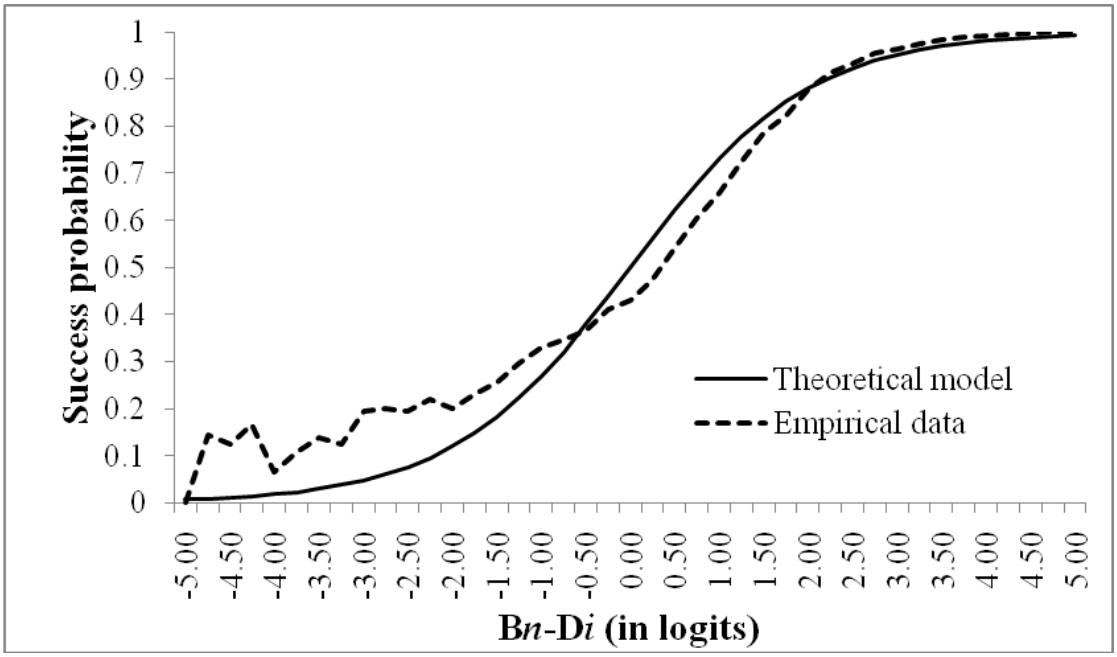


Figure 44. Success probability for the form section

person with lower ability meeting a more difficult item. The vertical axis shows the probability of a person with ability B_n succeeding on an item with difficulty D_i. The smooth line represents the theoretical model. The dotted line, which represents the empirical data obtained from the participants in Study 2, deviates from the expected model with smaller values of B_n-D_i. In other words, when people with low ability met difficult items, their success probabilities approached 25% (the expected percentage of correct responses by random guessing), which was higher than the model expectation.

Taken together, Figures 42-44 indicate that lucky guessing occurred when people with low ability met difficult items in the form section. The design of the present research may allow such random guessing to occur, because a) no *Don't know* options were provided, b) the participants had to choose one answer to go to the next item, and c) for validation purposes all the participants worked on items with varying levels of difficulty. Lucky guessing was corrected by deleting response records which have difficulty greater than $b + \ln(m\text{-}1)$, where b is the person's initial estimated ability and m is the number of options (Wright & Stone, 1979). As each item had four options, responses with an item difficulty greater than $b + 1.1$ were deleted. A total of 4,646 out of 87,620 (5.3%) responses were deleted as the result of this treatment.

Second, similar to the form section, outfit statistics and success probabilities were examined for the meaning section. Figure 45 illustrates the scatter plot of item difficulty and outfit *t* for this section. This figure indicates a tendency that difficult items are identified as misfitting. Figure 46 presents the scatter plot of person ability and outfit *t*. This figure indicates a weak tendency that low ability persons are identified as misfitting. Figure 47 illustrates the probability of success when a person with the ability B_n met an item with the difficulty D_i. This figure shows that the dotted line (empirical

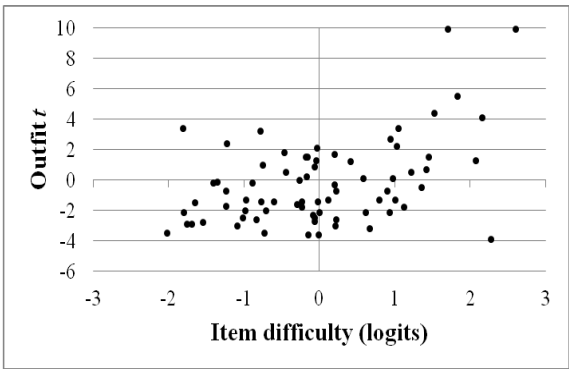


Figure 45. Item difficulty and outfit *t* for the meaning section

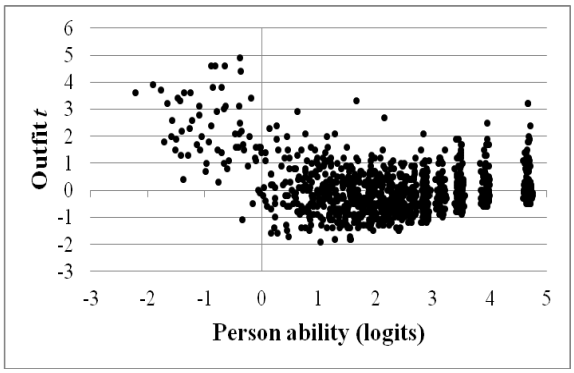


Figure 46. Person ability and outfit *t* for the meaning section

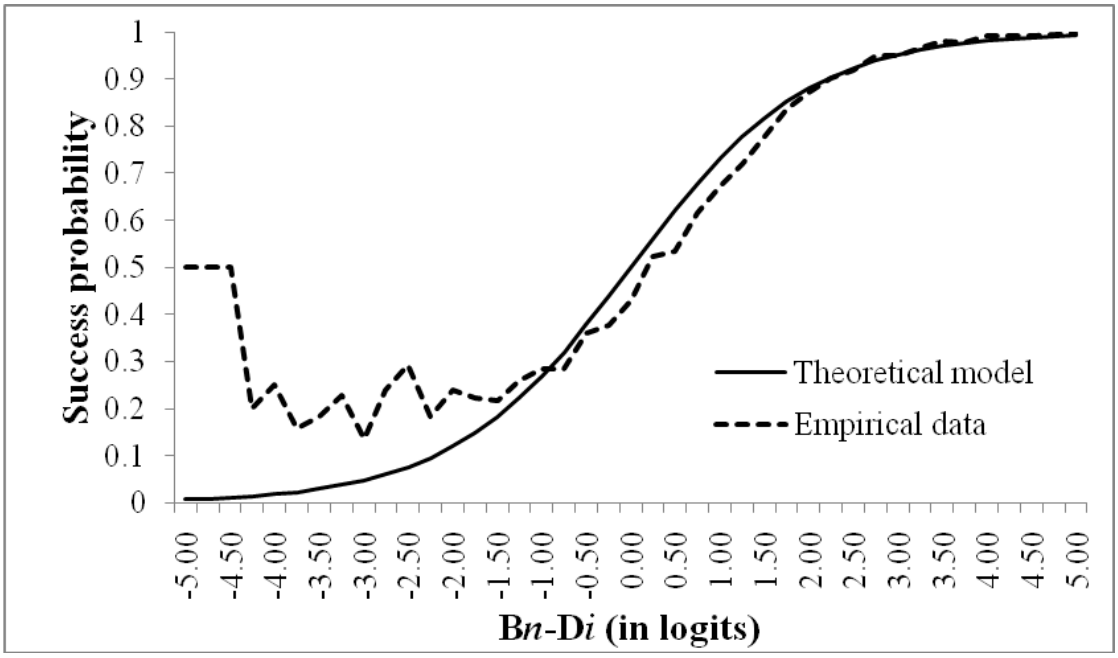


Figure 47. Success probability for the meaning section

data) deviates increasingly from the expected model with smaller values of B_n-D_i. In other words, when people with low ability met difficult items, their success probabilities approached 25% (the expected percentage of correct responses by random guessing), which was higher than the model expectation.[38] Taken together, Figures 45-47 indicate

[38] In Figure 47, the success probability was much higher (.5 at $-5.0<B_n-D_i<-4.5$) than the expectation because there were only six responses (three correct and three wrong responses).

that lucky guessing occurred when people with low ability met difficult items in the meaning section. As with the form section, lucky guessing was corrected by deleting response records with an item difficulty greater than b + 1.1. A total of 2,284 out of 72,792 (3.2%) responses were deleted as the result of this treatment.

Finally, outfit statistics and success probabilities were examined for the use section. Figure 48 illustrates the scatter plot of item difficulty and outfit t for this section. This figure indicates a tendency that difficult items are identified as misfitting. Figure 49 presents the scatter plot of person ability and outfit t. This figure does not clearly indicate a tendency that low ability persons are identified as misfit, but 84% of the misfit persons (48 out of 57) had person ability estimates below the average (1.11 logits). This may be taken as supportive evidence for the effect of lucky guessing. Figure 50 illustrates the probability of success when a person with the ability B_n met an item with the difficulty D_i. This figure shows that the dotted line (empirical data) deviates increasingly from the expected model with smaller values of B_n-D_i. In other words, when people with low ability met difficult items, their success probabilities approached 25% (the expected percentage of correct responses by random guessing), which was higher than the model expectation. Taken together, Figures 48-50 may indicate lucky guessing for responses with smaller B_n-D_i. As with the previous two sections, lucky guessing was corrected by deleting response records with an item difficulty greater than b + 1.1. A total of 6,076 out of 53,920 (11.3%) responses were deleted as the result of this treatment.

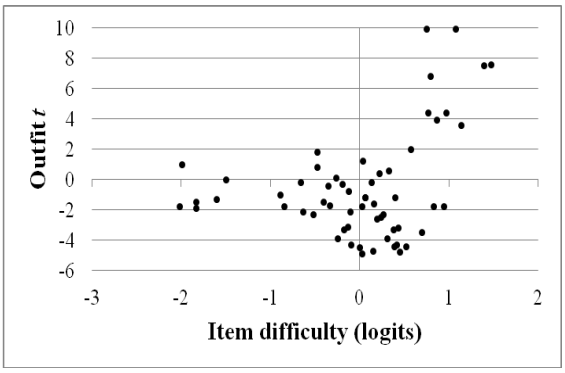


Figure 48. Item difficulty and outfit *t* for the use section

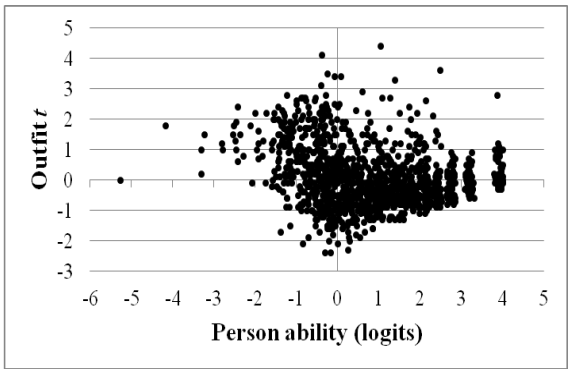


Figure 49. Person ability and outfit *t* for the use section

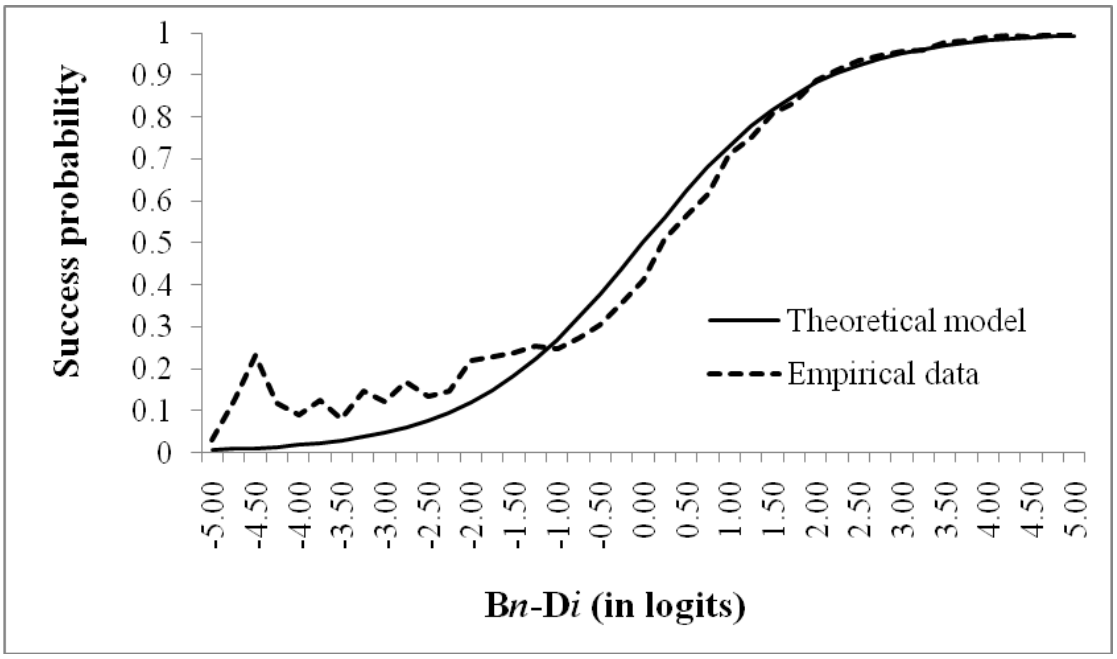


Figure 50. Success probability for the use section

In summary, lucky guessing was corrected for all three sections by deleting response records that have an item difficulty greater than $b + \ln(m\text{-}1)$ (Wright & Stone, 1979). The subsequent section discusses the validity of the revised WPT after the correction for lucky guessing.

6.2.4 Validity

This section aims to investigate the validity of the revised WPT. As with the validation of the GCT, the WPT was validated based on Messick's (1989, 1995) six aspects of construct validity (content, substantive, structural, generalizability, external, and consequential) and the two non-overlapping aspects (responsiveness and interpretability) proposed by the Medical Outcomes Trust Scientific Advisory Committee (1995) (see Section 4.6 for a detailed discussion). The subsequent sections provide evidence of the construct validity of the WPT from the eight aspects largely on the basis of Rasch measurement.

6.2.4.1 Content Aspect

The content aspect of construct validity was evaluated in terms of relevance, representativeness and technical quality (Messick, 1989, 1995). This section investigates the content aspect from each of the three sub-aspects.

Relevance

An in-depth discussion of the construct definition of word part knowledge and the tasks for measuring the construct was given in the previous chapter. Here are the key points.

- The present research focuses on written receptive knowledge of word parts. Word part knowledge involves 1) recognition of the form of the word part in a word, 2) knowing its meaning, and 3) knowing its use (part of speech) (Bauer & Nation, 1993; Nation, 2001; Tyler & Nagy, 1989):.
- The WPT had three sections (form, meaning, and use) in order to measure the three aspects of receptive word part knowledge.
- A word part, or an affix, was defined as a bound morph which co-occurs with bases which contain free morphs. The present research focused on derivational affixes instead of inflectional ones.

- The 118 word parts that were selected for the present research appeared in more than one word family in the first 10,000 word families in the BNC word lists. Allomorphs (word parts which vary in spelling or sound but not in meaning) were treated as different word parts.

- The quality of the 118 word parts was considered high, because these word parts covered a large proportion of the word parts that were listed or used in previous studies (Bauer & Nation, 1993; Bock, 1948; Carroll, 1940; Freyd & Baron, 1982; Harwood & Wright, 1956; Mochizuki, 1998; Mochizuki & Aizawa, 2000; Nagy, et al., 1993; Nation, 2001; Schmitt & Meara, 1997; Stauffer, 1942; Thorndike, 1941; Tyler & Nagy, 1989; Wysocki & Jenkins, 1987): an average of 67.4% for prefixes and 90.0% for suffixes (The low coverage for prefixes was mainly due to a different definition of affix).

- The test format for each section was determined by examining six aspects of test usefulness: reliability, construct, authenticity, interactiveness, impact, and practicality (Bachman & Palmer, 1996). The selected formats were considered to be most useful from these perspectives.

- All the selected word parts were measured in the form section.

- A total of 73 word parts with substantial meaning were measured in the meaning section. Word parts with highly abstract meaning such as *-ness* (state, condition, quality) in *happiness* and *-ment* (action, state, results) in *movement* were excluded.

- A total of 59 word parts that change the part of speech of the word base were included in the use section.

It should be reasonable to conclude that the test content is highly relevant to knowledge of English word parts because the tasks were created so that word part knowledge may be comprehensively measured in three different sections.

Representativeness

The WPT is considered to be highly representative of the construct domain, because 1) affixes were selected from the most frequent 10,000 word families which may be a minimum requirement for unassisted comprehension of written text (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006), 2) the selected affixes covered a large proportion of affixes that were identified or used in previous studies, and 3) all the

selected word parts were measured in the WPT.

Representativeness was also evaluated through the Rasch item difficulty hierarchy (Smith Jr., 2004b). The spread of item calibrations was examined by item strata. An item strata index indicates the number of statistically different levels of item difficulty, and is derived by the following formula:

$$\text{Item strata} = (4\,G_{item}+1)/3,$$

where G*item* is Rasch item separation. Item strata statistics need to be greater than 2.0 for useful tests (Smith Jr., 2004b, p. 106). The item strata statistics for the three sections are presented in Table 54. This table shows that each section had an item strata index of well above 2, which may be taken as supportive evidence for the representativeness of the tasks.

Table 54. Item strata for the three sections of the revised WPT

Section	Item strata
Form	13.91
Meaning	11.39
Use	13.25

Finally, representativeness was investigated by examining a Rasch person-item map to see whether there were gaps in the item difficulty hierarchy. Figure 51 is a person-item map for the form section. The far left of this figure shows a Rasch logit scale with the mean item difficulty being 0. In this figure, the item distribution is presented on the right. More difficult items are located towards the top and less difficult items are located towards the bottom. Figure 51 shows that there are no gaps in the item difficulty hierarchy between +2 and -2 logits, indicating a high degree of representativeness in terms of item difficulty for that range. It also shows that the

affix *-i* is by far the most difficult to recognise. A fit analysis identified this item as misfit (outfit $t = 4.9$, infit $t = 5.1$), perhaps because one of the distractors was chosen by a large number of people with high ability. This will be discussed later in this section.

Figure 52 is a person-item map for the meaning section. This figure shows that there are few gaps in the item difficulty hierarchy between +2.5 and -2.5 logits. Taken together with the sufficient number of statistically distinct levels (item strata = 11.39), the meaning section may be highly representative in terms of item difficulty.

Figure 53 is a person-item map for the use section. This figure shows that there are few gaps in the item difficulty hierarchy between +2 and -2 logits. Taken together with the sufficient number of statistically distinct levels (item strata = 13.25), the use section may be highly representative in terms of item difficulty.

	<More able persons>			<More difficult items>					
5	*###		+						
			|						
			|						
	#	T	|						
	#####		|		-i					
4			+							
	*####		|						
	*#		|						
	*#####		|						
	*###		|						
3	*####	S	|						
	*####		|						
	*####		|						
	*#######		|	T					
	*#######		|		-ster				
2	*##########		+			-let				
	*############		|		-atory	-esque	-ette			
	*############	M	|		-ar	-eer	-ery			
	##########		|		-th					
	*##########		|	S	-en -ite	-or -ory	-ways -ure	-fold		
1	*#########		+			-ance	-ancy	-dom			
	*########		|		-atic	-et	-ible			
	*#######		|		-an -ese	-ary -ion	-cy -most	-ee arch-	-ency	
	*######		|		-age	-ant	-ify	-ic	-ize	-ling
	*####	S	|		-al a-	-ence	-hood	-ent	-ian	-ity
0	*###		|	M	-ess	-ist	-ship	-ous	neo-	
	*#		|		-ism -ward	-some -wise	-ty mal-	-ise	-ive	
	*#		|		-ate	-ation	-er	circum-		
	*##		|		-ish	be-	il-	bi-	counter-	sur-
	*#		|		-ful	ab-	en-	-ition	-y	em-
-1	*		+			-less	-ness	co-	ir-		
	*#	T	|	S	-able	-ment	anti-	fore-	trans-	
	*#		|		im- super-	in-	pre-	mid-	mono-	post-
	*#		|		-ly non-	dis- semi-	ex-	auto-	de-	hyper-
	*		|		multi-	uni-				
-2	*		+			inter-	mis-	sub-	pro-	re-	un-
	*		|							
	*		|	T	micro-					
	*		|							
	*		|							
-3	*		+								
	<Less able persons>		|	<Less difficult items>						

Note. # = 8 persons; * = 1 to 7 persons; M = mean of the person or item estimates; S = one standard deviation from the mean; T = two standard deviations from the mean.

Figure 51. Person-item map for the form section

	<More able persons>			<More difficult items>					
4	#####	S	+						
			|						
	*###		|						
	*##		|						
	*##		|						
	##		|						
3	*#####		+						
	*##		|		-ette				
	*###		|						
	*###	M	|						
	*###		|	T	a-(not)	arch-			
	*###		|		in-				
2	*###		+			-et	-ling			
	*##		|		de-				
	*###		|						
	*#		|		-ship	counter-	neo-		
	*###		|		mal-				
	*##		|	S	-ite	-let	pro-	sur-	uni-
1	*#	S	+			-fold	ab-	ex-		
	*##		|		inter-				
	*		|		-dom	-ism	super-		
	*#		|		circum-				
	*		|		-less	sub-			
	*		|		co-	hyper-	mono-	non-	
0	*		+	M		-ful	-hood	-i	a-(toward)	bi-
						-il	im-	post-		
	*		|			-ible	-ster	-wise	dis-	fore-
						ir-	un-			
	*		|		-ward	auto-			
	*		|		-ent	-most			
	*	T	|		anti-				
	*		|		-able	-ant	-ary	micro-	semi-
-1	*		+			-or	multi-	trans-		
	*		|	S	-en	mis-	pre-		
	*		|		-er	-ess	re-		
	*		|		-ee	-ways			
	*		|		-an				
	*		|		-eer	mid-			
-2	*		+			-ian	-ist			
	*		|		-th				
	*		|	T	-ese				
	*		|						
	*		|						
	*		|						
-3	*		+							
	<Less able persons>			<Less difficult items>					

Note. # = 15 persons; * = 1 to 14 persons; M = mean of the person or item estimates; S = one standard deviation from the mean; T = two standard deviations from the mean.

Figure 52. Person-item map for the meaning section

	<More able persons>			<More difficulty items>				
3	#####		+					
	*########		|					
	*######	S	|					
	*#######		|					
	*########		|					
	*#######		|	-ways(adv)				
2	######		+	T	-wise(adv)				
	*#######		|		-ly(adj)				
	*#######		|					
	*#########		|		-atory(adj)				
	*########		|					
	*########		|		-ory(adj)	be-(v)			
1	#########	M	+	S	-ary(adj)	-ward(adv)	a-(adv)		
	#########		|		-y(n)	em-(v)			
	*######		|		-esque(adj)	-ty(n)			
	########		|		-en(v)	-ency(n)	en-(v)		
	#######		|		-age(n)	-ancy(n)	-ify(v)	-ise(v)	-ity(n)
						-ize(v)	-th(n)			
	*#########		|		-al(n)	-ant(adj)	-ar(adj)	-cy(n)	-some(adj)
0	*#######		+	M		-able(adj)	-ate(adj)	-ence(n)	-ery(n)	-less(adj)
						-ness(n)	-ure(n)			
	*##########		|		-en(adj)	-ible(adj)	-ic(adj)	-ion(n)	-ition(n)
	*#########		|		-ance(n)	-atic(adj)	-ation(n)		
	*########		|		-al(adj)	-ate(v)	-ive(adj)	-ly(adv)	-y(adj)
	*#########	S	|		-ful(adj)	-ment(n)			
	*########		|		-ent(adj)				
-1	*######		S	S	-ish(adj)	-ous(adj)			
	*#####		|						
	*##		|						
	*###		|		-ary(n)				
	###		|		-ant(n)				
	*#		|						
-2	*#		+	T	-ee(n)	-ent(n)			
	*#		|		-er(n)	-or(n)			
	*#		|						
	*	T	|						
	*		|						
	*		|						
-3	*#		+						
	<Less able persons>			<Less difficult items>				

Note. # = 5 persons; * = 1 to 4 persons; M = mean of the person or item estimates; S = one standard deviation from the mean; T = two standard deviations from the mean; n = noun; v = verb; adj = adjective; adv = adverb.

Figure 53. Person-item map for the use section

Technical Quality

Technical quality was investigated by examining the degree to which the empirical data fit the Rasch model (Smith Jr., 2004b). More specifically, point-measure correlations and fit statistics were investigated for each section. The present research used outfit and infit standardised *t* statistics as the primary criterion for detecting misfit items instead of outfit and infit unstandardised mean-square statistics, because the *t* statistics may identify a greater number of misfit items than mean-square statistics with a larger sample. Mean-square statistics allow an increasing number of items to be acceptable simply by increasing sample size (Smith, 2000; Smith, et al., 1998; Smith & Suh, 2003). With a sample size of more than 1,300, mean-square statistics might fail to identify a number of misfit items. It should be noted here that the *t* statistics might potentially identify good items as misfit with a large sample size (Karabatsos, 2000; Linacre, 2003; Smith, et al., 2008); thus, each misfit item (outfit $t > 2.0$ or infit $t > 2.0$) was carefully inspected to see whether it was really a bad item. Misfit items that met all the following four criteria (C1-C4) were regarded as being acceptable:

C1. Outfit and infit mean-square statistics were smaller than 1.5 which may be an indication of "useful" fit (Linacre, 2002, 2003, 2010a);

C2. The correct answer was higher in the average person ability than any of the three distractors;

C3. The correct answer showed a positive point-measure correlation, while the three distractors showed negative point-measure correlations; and

C4. The three distractors were not greatly different in the distribution of responses (the percentage of respondents who chose the distractor and their average person ability estimate).

First, the quality of the items was investigated for the form section. No items in this section showed negative or low positive point-measure correlations (less than .10), which supports the technical quality of this section. A subsequent fit analysis identified

ten items as misfit. Four of the ten misfit items had been identified as misfit in Study 1 (Table 55).

Table 55. Misfit items in the form section (Studies 1 & 2)

Item No.	Affix	Difficulty	S.E.	Outfit		Infit	
				t	MNSQ	*t*	MNSQ
68	-et	0.83	0.07	7.4	1.56	8.5	1.26
90	-ling	0.83	0.07	6.8	1.50	9.9	1.30
93	-most	1.03	0.07	4.4	1.28	6.3	1.18
105	-ways	1.13	0.07	2.4	1.13	4.9	1.13

Here are the details of the four misfit items.

- Item 68: *-et*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-et	-io	-ob	-ht
% chosen	68.2	12.0	8.4	11.4
Ave. ability (logits)	2.04	1.98	1.06	1.27
PT-measure corr.*	.20	.03	-.19	-.17

*point-measure correlation;

The analysis showed that Item 68 violated C1 (outfit mean-square = 1.56) and C3 (point-measure correlation of Distractor 1 = .03). A close look at Distractor 1 showed that it was chosen by a number of native English speakers with high ability; specifically, it was chosen by 27 native speakers with person ability estimates of larger than 3.03 logits (90% probability of succeeding on Item 68), while only seven non-native speakers with that range of ability estimates chose it. This may have been because they recalled infrequent words with the *-io* ending such as *cheerio* and *mustachio* which might be mistaken as *cheer* + *-io* and *mustache* + *-io*. The mean-square fit statistics got acceptably improved (<1.5) if the 27 native speakers were removed from the analysis (outfit mean-square = 1.24; infit mean-square = 1.20). The statistics of Distractor 1 also

became acceptable with this treatment (% chosen = 10.0%, average ability = 1.61; point-measure correlation = -.05). Given that the WPT is designed primarily for English learners, it should be reasonable to conclude that this item is acceptable.

- Item 90: *-ling*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-ling	-reat	-tute	-bute
% chosen	68.2	7.6	12.4	11.8
Ave. ability (logits)	2.02	0.91	1.83	1.59
PT-measure corr.	.18	-.21	-.01	-.08

Item 90 may be acceptable because it meets the four criteria C1-C4.

- Item 93: *-most*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-most	-oard	-ogue	-laim
% chosen	65.4	8.1	17.3	9.2
Ave. ability (logits)	2.16	1.04	1.75	1.15
PT-measure corr.	.28	-.21	-.06	-.19

Item 93 may be acceptable because it meets the four criteria C1-C4. Distractor 2 was chosen by a relatively large number of persons with high ability, but it was unlikely to be mistaken for a real affix.

- Item 105: *-ways*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-ways	-ause	-ript	-oice
% chosen	63.9	16.4	7.2	12.5
Ave. ability (logits)	2.22	1.47	1.41	1.27
PT-measure corr.	.32	-.16	-.12	-.20

Item 105 may be acceptable because it meets the four criteria C1-C4.

Rasch analysis in Study 2 detected six misfit items that had been acceptable in Study 1 (Table 56).

Table 56. Misfit items in the form section (Study 2 only)

Item No.	Affix	Difficulty	S.E.	Outfit		Infit	
				t	MNSQ	*t*	MNSQ
101	-th	1.33	0.09	7.2	1.58	8.1	1.31
73	-i	4.20	0.11	4.9	1.38	5.1	1.31
100	-ster	2.23	0.09	3.5	1.19	3.0	1.11
103	-ure	1.52	0.09	2.8	1.20	5.1	1.19
45	-an	0.57	0.10	2.6	1.29	3.8	1.17
89	-let	2.02	0.09	2.2	1.13	4.2	1.15

Here are the six misfit items.

- Item 101: *-th*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-th	-ak	-wn	-ul
% chosen	61.1	6.0	6.0	26.9
Ave. ability (logits)	2.10	1.26	1.49	1.87
PT-measure corr.	.16	-.15	-.10	-.04
Frequency in 10k wds	95	25	41	128
Replaced by				**-ol**
Frequency in 10k wds				**29**

The analysis showed that Item 101 violated C1 (outfit mean-square value = 1.58). A close look at the distractors showed that Distractor 3 was chosen by a large number of persons (26.9%) with relatively high ability (1.87). This distractor may have been popular, because there are a large number of words with the *-ul* ending many of which were part of the *-ful* ending (e.g., *beautiful* and *careful*). To avoid this confusion, this distractor was replaced by *-ol* (e.g., *alcohol* and *school*) which is less frequent than *-ul* and is less likely to be perceived to be a real suffix.

- Item 73: *-i*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-i	-u	-r	-p
% chosen	30.7	3.7	62.4	3.2
Ave. ability (logits)	3.44	2.78	3.25	2.52
PT-measure corr.	.12	-.10	-.02	-.15
Frequency in 10k wds	9	17	2,415	287
Replaced by			**-w**	
Frequency in 10k wds			**133**	

The analysis showed that Item 73 violated C4 (Distractor 2 was chosen by a large number of persons (62.4%) with relatively high ability (3.25)). Distractor 2 may have been popular, because there are a large number of words with the *-r* ending many of which were part of the *-er/-or* ending (e.g., *teacher* and *actor*). Some of the words might have been mistakenly divided into a word base + *-r* (e.g., *maker* = *make* + *-r*, and *baker* = *bake* + *-r*). To avoid this confusion, this distractor was replaced by *-w* (e.g., *follow* and *window*) which is less frequent than *-r* and is less likely to be viewed as a real suffix.

- Item 100: *-ster*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-ster	-ange	-ulum	-unch
% chosen	49.0	33.9	10.2	6.9
Ave. ability (logits)	2.58	1.94	2.18	1.52
PT-measure corr.	.31	-.21	-.02	-.19

Item 100 may be acceptable because it meets the four criteria C1-C4. Distractor 1 was chosen by a relatively large number of persons, but their average person ability was not typically high.

- Item 103: *-ure*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-ure	-mph	-oke	-nse
% chosen	58.1	9.1	8.1	24.7
Ave. ability (logits)	2.30	1.53	1.47	1.66
PT-measure corr.	.29	-.12	-.13	-.17

Item 103 may be acceptable because it meets the four criteria C1-C4. Distractor 3 was chosen by a relatively large number of persons, but their average person ability was not extremely high compared to the other distractors.

- Item 45: *-an*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-an	-oo	-ue	-lc
% chosen	72.2	9.5	15.4	2.9
Ave. ability (logits)	2.09	1.49	1.18	0.93
PT-measure corr.	.28	-.09	-.22	-.12

Item 45 may be acceptable because it meets the four criteria C1-C4.

- Item 89: *-let*

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	-let	-que	-nct	-uid
% chosen	51.8	33.9	5.3	9.0
Ave. ability (logits)	2.53	1.90	1.73	1.54
PT-measure corr.	.32	-.18	-.09	-.18

Item 89 may be acceptable because it meets the four criteria C1-C4. Distractor 1 was chosen by a relatively large number of persons, but their average person ability was not extremely high compared to the other distractors.

Second, the quality of the items in the meaning section was investigated. No items in this section showed negative or low positive point-measure correlations (less than .10), which supports the technical quality of this section. A subsequent fit analysis

identified nine items as misfit. Two of the nine misfit items had been identified as misfit in Study 1 (Table 57).

Table 57. Misfit items in the meaning section (Studies 1 & 2)

Item No.	Affix	Difficulty	S.E.	Outfit		Infit	
				t	MNSQ	*t*	MNSQ
65	-ling	1.97	0.07	9.9	2.00	9.9	1.47
54	-fold	1.04	0.08	2.5	1.25	3.6	1.16

Here are the details of the two misfit items.

- Item 65: *-ling* (weakling; underling)

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	connected with	opposite	too much	direction
% chosen	70.9	5.4	14.0	9.7
Ave. ability (logits)	3.57	2.05	3.07	2.57
PT-measure corr.	.24	-.19	-.06	-.15
Replaced by			**together**	

The analysis showed that Item 65 violated C1 (outfit mean-square = 2.00). A close look at the distractors revealed that Distractor 2 was chosen by a relatively large number of persons (14.0%) with high ability (3.07). This may have been because this distractor (*too much*) was too close to the correct meaning. The test-takers may have thought a weakling to be *too* weak a person and an underling to be a person in *too* low a position. This distractor was replaced by *together* which may be further away from the correct meaning.

- Item 54: *-fold* (twofold; threefold)

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	times	under	self	over
% chosen	81.0	4.3	3.1	11.6
Ave. ability (logits)	3.50	1.76	1.63	2.02
PT-measure corr.	.37	-.18	-.17	-.25

Item 54 may be acceptable because it meets the four criteria C1-C4.

Rasch analysis in Study 2 detected six misfit items that had been acceptable in Study 1 (Table 58).

Table 58. Misfit items in the meaning section (Study 2 only)

Item No.	Affix	Difficulty	S.E.	Outfit		Infit	
				t	MNSQ	*t*	MNSQ
70	-th	-2.09	0.26	4.6	4.43	-1.1	0.79
43	-ary	-0.88	0.17	3.6	2.76	2.2	1.28
32	re-	-1.35	0.20	3.1	2.55	1.2	1.19
55	-ful	0.00	0.14	2.4	1.59	1.9	1.18
66	-most	-0.49	0.11	2.3	1.53	0.1	1.00
59	-ible	-0.20	0.11	2.1	1.41	0.2	1.01

Here are the details of the six misfit items.

- Item 70: *-th* (fourth; sixth)

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	number	person	not	small
% chosen	96.8	1.2	0.4	1.6
Ave. ability (logits)	3.16	0.66	-1.49	-1.95
PT-measure corr.	.37	-.13	-.15	-.32

The analysis showed that Item 70 violated C1 (outfit mean-square = 4.43), and Distractor 1 was chosen by persons with relatively high ability (0.66). A close look at this distractor revealed that it was chosen by two native speakers with high ability estimates (4.8 logits = 99.9% probability of succeeding on this item). This may have been due to careless mistakes which may have occurred because the test-takers were not allowed to go back to the previous items to change the answers once they clicked on one option. The fit statistics got acceptably improved if the two native speakers were removed from the analysis (outfit mean-square = 1.21; infit mean-square = 0.74). The statistics of Distractor 1 also became acceptable with this treatment (% chosen = 0.9%,

average ability = -0.72; point-measure correlation = -.18). Thus, this item was considered acceptable.

- Item 43: *-ary* (secretary; commentary)

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	person/thing	away from	after	many
% chosen	92.9	2.2	2.2	2.7
Ave. ability (logits)	3.14	0.49	1.94	0.79
PT-measure corr.	.28	-.20	-.08	-.19

The analysis showed that Item 43 violated C1 (outfit mean-square = 2.76), and Distractor 2 was chosen by persons with relatively high ability (1.94). A close look at this distractor revealed that it was chosen by six persons with high ability estimates (larger than 3.72 logits = more than 99% probability of succeeding on this item). This may have been due to careless mistakes. The fit statistics got acceptably improved if the six persons were removed from the analysis (outfit mean-square = 1.41; infit mean-square = 1.22). The statistics of Distractor 2 also became acceptable with this treatment (% chosen = 1.4%, average ability = 0.65; point-measure correlation = -.15). Thus, this item was considered acceptable.

- Item 32: *re-* (replay; rebuild)

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	again	person	before	female
% chosen	95.1	0.7	3.4	0.8
Ave. ability (logits)	3.08	-0.53	1.45	-1.46
PT-measure corr.	.28	-.16	-.15	-.22

The analysis showed that Item 32 violated C1 (outfit mean-square = 2.55), and Distractor 2 was chosen by a relatively large number of persons (3.4%) with relatively high ability (1.45). A close look at this distractor revealed that it was chosen by four persons with high ability estimates (larger than 3.25 logits = more than 99% probability

of succeeding on this item). This may have been due to careless mistakes. The fit statistics got acceptably improved if the four persons were removed from the analysis (outfit mean-square = 0.97; infit mean-square = 1.14). The statistics of Distractor 2 also became acceptable with this treatment (% chosen = 2.8%, average ability = 0.91; point-measure correlation = -.19). Thus, this item was considered acceptable.

- Item 55: *-ful* (handful; mouthful)

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	amount	small	not	person
% chosen	88.7	3.1	1.4	6.8
Ave. ability (logits)	3.23	1.40	0.39	1.52
PT-measure corr.	.33	-.16	-.18	-.23

The analysis showed that Item 55 violated C1 (outfit mean-square = 1.59), and Distractor 3 was chosen by a relatively large number of persons (6.8%) with relatively high ability (1.52). A close look at this distractor revealed that it was chosen by two persons with high ability estimates (larger than 4.60 logits = more than 99% probability of succeeding on this item). This may have been due to careless mistakes. The fit statistics got acceptably improved if the two persons were removed from the analysis (outfit mean-square = 1.34; infit mean-square = 1.16). The statistics of Distractor 3 also became acceptable with this treatment (% chosen = 6.5%, average ability = 1.40; point-measure correlation = -.24). Thus, this item was considered acceptable.

- Item 66: *-most* (topmost; uppermost)

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	the furthest	person	opposite	half
% chosen	91.3	3.1	3.4	2.3
Ave. ability (logits)	3.25	0.45	0.34	0.95
PT-measure corr.	.41	-.24	-.26	-.17

Item 66 may be acceptable because it meets the four criteria C1-C4.

- Item 59: *-ible* (accessible; convertible)

	Correct	Dstractor 1	Distractor 2	Distractor 3
Option	can be	person/place	after	times
% chosen	90.1	4.3	2.3	3.4
Ave. ability (logits)	3.29	1.41	0.38	0.59
PT-measure corr.	.39	-.19	-.22	-.25

Item 59 may be acceptable because it meets the four criteria C1-C4.

Finally, the quality of the items in the use section was investigated. No items in this section showed negative or low positive point-measure correlations (less than .10), which supports the technical quality of this section. A subsequent fit analysis identified eight items as misfit. Four of the eight misfit items had been identified as misfit in Study 1 (Table 59).

Table 59. Misfit items in the use section (Studies 1 & 2)

Item No.	Affix	Difficulty	S.E.	Outfit		Infit	
				t	MNSQ	*t*	MNSQ
44	-ly(adj)	1.89	0.08	9.9	1.79	9.9	1.57
58	-y(adj)	-0.58	0.08	3.3	1.40	4.7	1.18
14	-ary(adj)	1.00	0.07	2.5	1.15	3.4	1.11
16	-ate(adj)	0.02	0.07	2.1	1.18	5.7	1.21

The correct answer for all four items in Table 59 was *adjective*. Here are the details of the four items.

- Item 44: *-ly* (adjective) (lively; friendly)

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	51.0	0.4	1.2	47.4
Ave. ability (logits)	2.26	-0.38	0.96	1.79
PT-measure corr.	.18	-.11	-.08	-.15

The analysis showed that Item 44 violated C1 (outfit mean-square = 1.79, infit mean-square =1.57) and C4 (Distractor 3 was chosen by a large number of persons with high ability). This may be because the majority of words with the *-ly* ending are adverbs (e.g., *widely*). As discussed in Section 6.1.4.3, however, the two example words (*lively* and *friendly*) are high-frequency words, and are used only as adjectives; thus, this item does not need to be rewritten. No change was made to this item but it needs watching for future use of the test.

- Item 58: *-y* (adjective) (lucky; healthy)

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	75.3	7.5	3.9	13.3
Ave. ability (logits)	1.60	0.04	-1.25	0.08
PT-measure corr.	.41	-.18	-.27	-.23

- Item 14:*-ary* (adjective) (revolutionary; parliamentary)

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	62.0	24.0	1.4	12.6
Ave. ability (logits)	2.29	1.05	0.08	0.71
PT-measure corr.	.44	-.26	-.13	-.26

- Item 16: *-ate* (adjective) (passionate; fortunate)

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adjective	noun	verb	adverb
% chosen	69.9	9.7	10.5	9.9
Ave. ability (logits)	1.82	0.34	0.51	-0.05
PT-measure corr.	.42	-.19	-.17	-.27

Items 58, 14, and 16 may be acceptable because they meet the four criteria C1-C4.

Rasch analysis in Study 2 detected four misfit items that had been acceptable in Study 1 (Table 60).

Table 60. Misfit items in the use section (Study 2 only)

Item No.	Affix	Difficulty	S.E.	Outfit		Infit	
				t	MNSQ	*t*	MNSQ
1	a-(adv)	1.00	0.07	9.8	1.68	9.9	1.38
55	-ward(adv)	1.03	0.10	3.9	1.34	3.8	1.18
57	-wise(adv)	2.07	0.12	3.9	1.26	4.2	1.21
56	-ways(adv)	2.14	0.11	2.2	1.14	2.7	1.13

The correct answer for these four items in Table 60 was *adverb*. Here are the details of the four items.

- Item 1: *a-* (adverb) (ahead; aside)

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adverb	noun	verb	adjective
% chosen	61.5	10.8	2.8	24.8
Ave. ability (logits)	2.09	0.82	0.35	1.45
PT-measure corr.	.29	-.21	-.16	-.11

The analysis showed that Item 1 violated C1 (outfit mean-square = 1.68), and Distractor 3 (adjective) was chosen by many people (24.8%) with relatively high ability (1.45). This may have been because some words with the prefix *a-* could also be used as an adjective (e.g., *asleep*). However, the two example words were high-frequency words (2nd 1,000 for *ahead* and 3rd 1,000 for *aside* in the BNC word lists) and had no adjective usages in the BNC; thus, there were no better alternatives to rewrite. No change was made to this item but it needs watching for future use of the test.

- Item 55: *-ward* (adverb) (upward; backward)

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adverb	noun	verb	adjective
% chosen	61.5	8.9	4.7	24.9
Ave. ability (logits)	2.22	0.83	0.48	1.23
PT-measure corr.	.38	-.20	-.19	-.21

- Item 57: -*wise* (adverb) (clockwise; stepwise)

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adverb	noun	verb	adjective
% chosen	51.2	12.8	2.5	33.5
Ave. ability (logits)	2.66	1.15	1.16	1.98
PT-measure corr.	.36	-.31	-.13	-.12

- Item 56: -*ways* (adverb) (sideways; lengthways)

	Correct	Distractor 1	Distractor 2	Distractor 3
Option	adverb	noun	verb	adjective
% chosen	49.6	19.6	1.0	29.8
Ave. ability (logits)	2.73	1.52	1.65	1.79
PT-measure corr.	.41	-.32	-.04	-.14

The analyses showed that Items 55, 57, and 56 met C1-C3, but Distractor 3 (adjective) was chosen by many people with relatively high ability. This may be because the suffixes -*ward*, -*wise*, and -*ways* can also make an adjective (e.g., *an upward trend*, *a clockwise direction*, and *a sideways glance*). Although these three items had been acceptable both in a pilot study with ten highly proficient native and non-native English speakers (see Section 5.5.4.3) and in Study 1 with Japanese learners of English, they were excluded from the use section, because they had two possible correct answers (adverb and adjective) and these three suffixes were measured in the form and the meaning sections.

This subsection has looked at the content aspect of construct validity. The item analysis based on the Rasch model revealed that the vast majority of the items were acceptable, which may be taken as empirical support for the content aspect of the construct validity of the WPT. Table 61 summarises the eight unacceptable items and how these items were treated. The top five items in the table (-*th*, and -*i* for the form section, -*ling* for the meaning section, and -*ly* and *a*- for the use section) need to be

inspected for future use of the WPT.

Table 61. Unacceptable items and their remedy

Section	Affix	Remedy
Form	-th	Rewritten
	-i	Rewritten
Meaning	-ling	Rewritten
Use	-ly (adjective)	None
	a- (adverb)	None
	-ward	Omitted
	-ways	Omitted
	-wise	Omitted

6.2.4.2 Substantive Aspect

The substantive aspect of construct validity was evaluated by examining whether the empirical item hierarchy was presented as predicted by theoretical argument and whether each person's response pattern was consistent with that item hierarchy (Smith Jr., 2004b). To begin with, the relationship between theoretical and empirical item hierarchy was examined for each of the three sections.

It was hypothesised that the difficulty order of the WPT items would partly be determined by the frequency of affixes, or the number of words in which an affix occurs, because learners would have a greater chance of meeting and learning an affix that occurs in a greater number of words in authentic context. Given that higher frequency words tend to be better known than lower frequency words (Beglar, 2010; Schmitt, et al., 2001), it would be reasonable to predict that an affix that occurs in a greater number of words is more likely to be answered correctly. In fact, research with a lexical decision task has indicated that the frequency of affixes and bases has an effect on the recognition of affixed words (Bradley, 1979; Cole, Beauvillain, & Segui, 1989; Taft, 1979). This hypothesis was examined by investigating the correlation between the item

difficulty estimate and the frequency of the affix. The affix frequency was calculated in two ways based on the two different units of counting a word: the number of lemmas (counting inflected forms as one word; e.g., *disagree*, *disagrees*, *disagreed*, and *disagreeing* are counted as one) and the number of tokens (the summed number of occurrences of words containing the affix) in the BNC. Lemma was used as a unit of counting a word, because the present research focuses on derivational affixes, and not on inflectional affixes which typically result in the same meaning and the same syntactic property as the word base (see Section 5.2 for details). Type (the number of different word forms) was not used, because the frequency is affected by whether the affixed words can be inflected or not. For example, the suffix *-ize* makes a verb and typically produces four word types per word base by inflection (e.g., *generalize*, *generalizes*, *generalized*, and *generalizing* for the base *general*), while words with the suffix *-some* are adjectives and have no inflected forms (e.g., *troublesome*). Word family (counting inflectional and some derivational forms as one word) was not used either, because the present research aims to investigate learners' overall knowledge of derivational affixes without any presupposition that some affixes are so frequent and transparent in meaning that words with those affixes may be considered as members of a word family. Table 62 presents Spearman's rank correlation coefficients between the Rasch item difficulty and the affix frequency as calculated by token and lemma for each of the three sections. Negative correlations were expected because a higher-frequency affix (larger number in frequency) would be less difficult (smaller number in item difficulty).

Table 62. Correlation coefficients between the item difficulty estimates and the affix frequency for the three sections.

Form section (*n*=107)		Meaning section (*n*=73)		Use section (*n*=56)	
Lemma	Token	Lemma	Token	Lemma	Token
-.487*	.030	-.200	-.398*	-.515*	-.527*

Note: $p<.05$.

Given that affix knowledge may be affected by productivity, regularity, instruction, L1 knowledge, as well as frequency (Bauer & Nation, 1993; Mochizuki & Aizawa, 2000; Tyler & Nagy, 1989), four of the significant correlation coefficients in Table 62 (ρ=-.487, -.398, -.515, and -.527) may be considered acceptably high. For the form section, no significant correlation was found between the token frequency and the item difficulty. This may be because the prefix items tended to be easier than the suffix items although suffixed words were much more frequent than prefixed words (Table 63).

Table 63. Means, standard deviations, *t*-statistics, and effect sizes of the item difficulty and the frequency between prefixes and suffixes for the form section

		Mean	*S.D.*	*t*	*d.f.*	*p*	*r*
Difficulty	Prefix	-0.97	0.74	8.87	105	.000	.67
	Suffix	0.60	0.97				
Frequency (token)	Prefix	14,747	25,952	3.95	67.4*	.000	.33
	Suffix	132,244	239,573				

*The degree of freedom was corrected because Levene's test revealed that the null hypothesis of the homogeneity of variance was rejected (F=20.02, p=.000).

The recognition of suffixes was more difficult than the recognition of prefixes, perhaps because the test-takers were presented with only suffixes without the beginning of any word which may be the most salient part in word recognition (Cutler, Hawkins, & Gilligan, 1985; Taft, 1985; Taft & Forster, 1976). Research (Cole, et al., 1989; Segui & Zubizaretta, 1985) has also shown that word frequency has differential effects on the recognition of prefixed and suffixed words, indicating differential lexical access procedures for these two types of words. Based on these research findings, the

correlations between the item difficulty and the token frequency were analysed separately for the prefix and the suffix items. The results showed that the correlations were acceptably high both for the prefix items (ρ = -.331, p < .05) and the suffix items (ρ = -.494, p < .05). This may be taken as supportive evidence for the substantive aspect of construct validity of the WPT.

Table 62 also showed that no significant correlations were found between the lemma frequency and the item difficulty for the meaning section (ρ = -.200). This may have been because three of the affixes (*-an, -ese,* and *-i*) typically attach to a small number of names of countries and places (smaller number in lemma) but these affixes were relatively easy (smaller number in difficulty) (see Table 64). A significant correlation coefficient of ρ = -.244 (p < .05) was derived with the exclusion of these three items. Taken together, empirical evidence may support the hypothesis that the difficulty order of the WPT items would be partly determined by the frequency of the affixes.

Table 64. Relatively easy affixes with low frequency for the meaning section

Affix	Example words	Lemma	Difficulty
-an	American, European	15	-1.73
-ese	Japanese Vietnamese	10	-2.28
-i	Israeli Iraqi	9	-0.08
Average of overall items		153	0.00

Other factors that may affect the difficulty order of the WPT items are productivity (the likelihood that the affix is used to make new words), predictability (the degree of predictability of the meaning of the affix), and regularity (the degree of change in the form of the word base when the affix is attached) (Bauer & Nation, 1993; Tyler & Nagy,

1989). However, it was hypothesised that these factors would have little influence on the difficulty order of the WPT items, because 1) productivity may be related more closely to productive affix knowledge which involves knowledge of allowable affixes for bases than to receptive affix knowledge which is the focus of the present research (see Section 5.1), 2) the meanings of the affixes may be predictable on the WPT because it deals with the most frequent meaning of each affix (see Section 5.5.3.3), and 3) two example words were selected for each affix from the most regularly affixed words (see Section 5.5.3.2). This hypothesis was examined using Bauer and Nation's (1993) seven levels of affixes which were determined based on the following criteria: frequency, productivity, predictability, and regularity. Here is the description of Levels 3 to 7 (Levels 1 and 2 are omitted here because they deal with word types and inflectional suffixes which are not the focus of the present research).

Level 3: The most frequent and regular derivational affixes,
Level 4: Frequent, orthographically regular affixes,
Level 5: Regular but infrequent affixes,
Level 6: Frequent but irregular affixes, and
Level 7: Classical roots and affixes.

The hypothesis predicted that the WPT items would show a difficulty order of affixes at Levels 3, 4, 6, and 5 with Level 3 being the easiest, because productivity, predictability, and regularity would not significantly affect the item difficulty while frequency would. No prediction was possible for affixes at Level 7, because no information was provided on frequency, productivity, predictability, or regularity. Figures 54-56 present the mean item difficulty and the 95% confidence interval for the four levels of Bauer and Nation's affix list (Levels 3-6) for the three sections. These figures show that the mean item

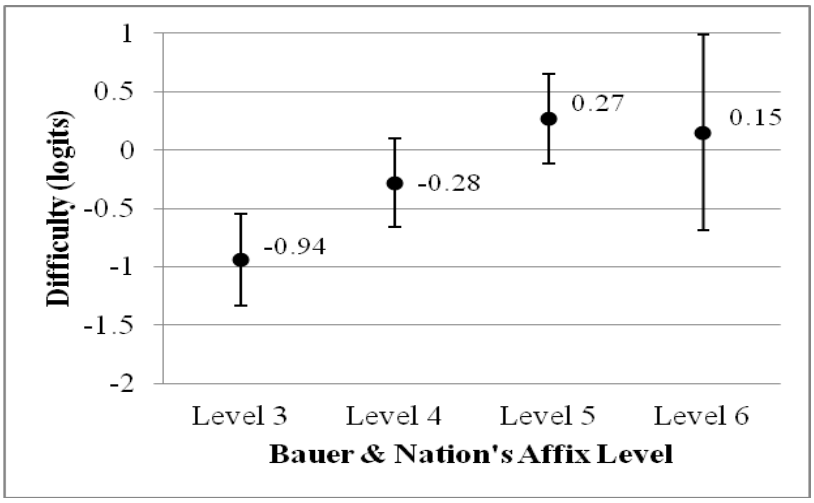


Figure 54. Mean item difficulty and 95% confidence interval according to Bauer and Nation's affix level for the form section

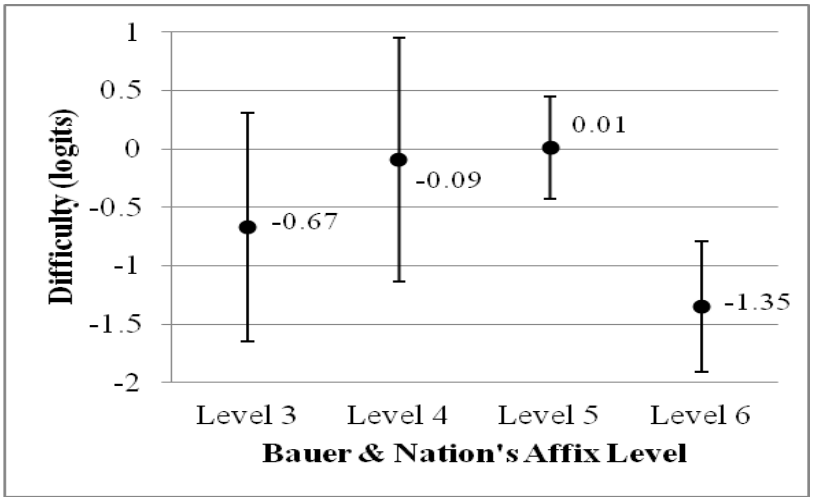


Figure 55. Mean item difficulty and 95% confidence interval according to Bauer and Nation's affix level for the meaning section

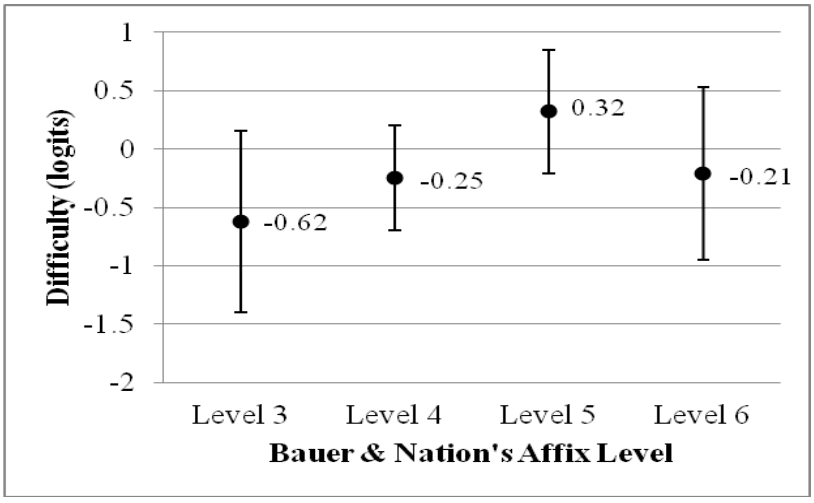


Figure 56. Mean item difficulty and 95% confidence interval according to Bauer and Nation's affix level for the use section

difficulties were ordered as predicted except for Level 6 of the meaning section,[39] although one-way ANOVAs showed that no significant difference was found for the meaning ($F(3,49)=1.52$, $p=.221$) and the use sections ($F(3,41)=1.94$, $p=.138$). A one-way ANOVA detected a significant difference for the form section ($F(3,74)=3.12$, $p=.031$). Tukey post-hoc comparisons of the four levels indicated that Level 3 affixes ($M=-0.94$, 95%CI[-1.34, -0.55]) were significantly easier than Level 5 affixes ($M=0.27$, 95%CI[-0.11, 0.66]), $p=.037$, but comparisons between the other levels were not statistically significant at $p<.05$. This indicates a weak tendency that the difficulties of the WPT items were ordered as predicted: Levels 3, 4, 6, and 5 with Level 3 being the least difficult. This may be taken as supportive evidence for the hypothesis that productivity, predictability, and regularity would have little effect on the WPT item difficulty, while frequency would.

The substantive aspect of construct validity was also evaluated by examining the consistency of each person's response pattern with the item hierarchy. More specifically, Rasch person fit statistics were calculated for each section. A misfit person was defined as having the person fit statistic of outfit $t > 2.0$ or infit $t > 2.0$ (underfit), or outfit $t < -2.0$ or infit $t < -2.0$ (overfit). Table 65 presents the number of misfit persons for each section. Each section had the misfit rate of less than 5% which was expected to occur by chance given the nature of the *z* distribution. This indicates that the test-takers' response pattern corresponded to the modelled difficulty order, which may be taken as supportive evidence for the substantive aspect of construct validity.

[39] There were only three affixes at Level 6 of the meaning section: *pre-* (*preschool*), *re-* (*replay*), and *-ee* (*employee*). The other seven affixes at Level 6 were not measured in the meaning section but in the form and the use sections. Two of those three affixes attach to a large number of different bases (548 lemmas for *pre-* and 561 lemmas for *re-*) compared to the average of all the 118 affixes (153 lemmas), which may be the reason for the low average item difficulty estimate.

Table 65. Number of misfit persons

Section	Number of underfit persons	Number of overfit persons	Total	%
Form	59	3	62	4.6
Meaning	22	1	23	1.7
Use	51	3	54	4.0

This subsection provided evidence for the substantive aspect of construct validity. Each of the three sections had the item difficulty order and the person ability order that were hypothesised by theoretical argument.

6.2.4.3 Structural Aspect

The structural aspect of construct validity was evaluated by examining the unidimensionality of the test (see Section 4.6.3 for details). Unidimensionality was examined by 1) item correlations, 2) fit statistics, and 3) principal components analysis (PCA) of standardised residuals without rotation (Linacre, 1995). Item correlations were investigated by computing the point-measure correlation. The results showed that no items had an unacceptably low point-measure correlation of smaller than .10. This indicates that unidimensionality holds in terms of item correlation.

A second way of investigating dimensionality was to identify misfit items (outfit $t > 2.0$ or infit $t > 2.0$). A close look at the misfit items indicated that eight items (two for the form section, one for the meaning section, and five for the use section) were unacceptable. These unacceptable items might degrade unidimensionality, but six of these items were rewritten or omitted from the analysis (see Table 61). This indicates that the WPT items may largely conform to the model which requires that measures be unidimensional. This may be taken as evidence for unidimensionality.

Finally, the PCA of standardised residuals was performed for each section in order to examine whether there was only a small amount of variance in the residuals accounted for by other components (dimensions) than the Rasch model which extracts the first major component in the observations. The scores generated by the three sections were regarded as unidimensional if the data met the following criteria: 1) the first contrast (largest secondary component) had an eigenvalue (standardised residual variance) of less than 3 (Linacre, 2010a, p. 444; Linacre & Tennant, 2009), and 2) the eigenvalues of other contrasts reached an asymptote at the first contrast (Stevens, 2002; Wolfe & Smith Jr., 2007). In order to investigate this, the scree plot for each section was examined (Figures 57-59). Figure 57 presents the scree plot for the form section. This figure shows that the first contrast had an acceptable eigenvalue (2.6) but had a larger eigenvalue than the second (2.0) to the fifth contrast (1.6), indicating that the data might have a secondary dimension. In order to further examine this, the contrast between strongly positively loading items and strongly negatively loading items on the first contrast was investigated to see whether they were substantively different enough to deserve the construction of two separate subtests (Linacre, 2010a, p. 445). Table 66 presents the ten items with the largest positive and negative loadings on the secondary dimension (first contrast). This table shows that prefix items had positive loadings on the first contrast, while suffix items had negative loadings on it, indicating that the prefix items and the suffix items might be measuring different constructs. However, it may not be effective to split the items of the form section into two subtests (prefix form section and suffix form section), because the person ability estimates produced by the prefix items were highly consistent with those produced by the suffix items (Pearson's *r* = .719, and .908 with correction for attenuation due to measurement error

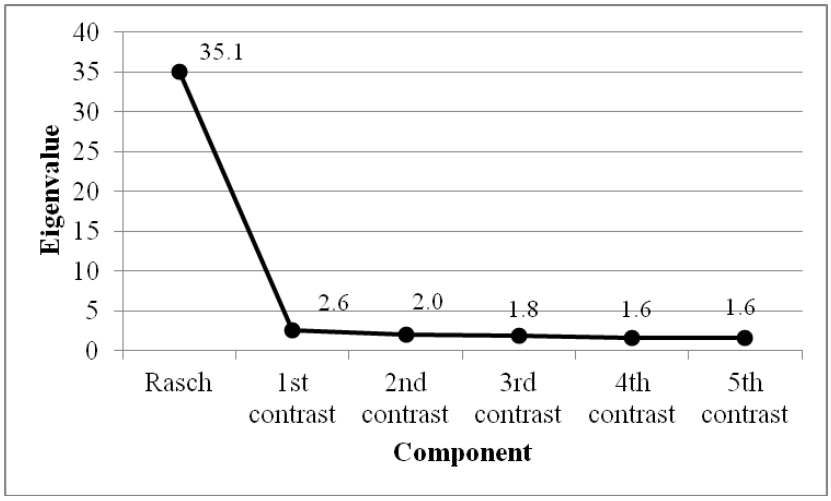


Figure 57. Scree plot for the form section

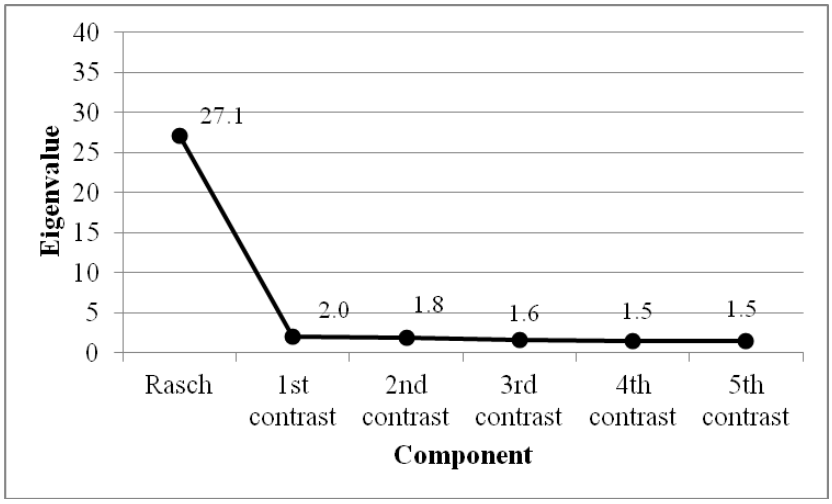


Figure 58. Scree plot for the meaning section

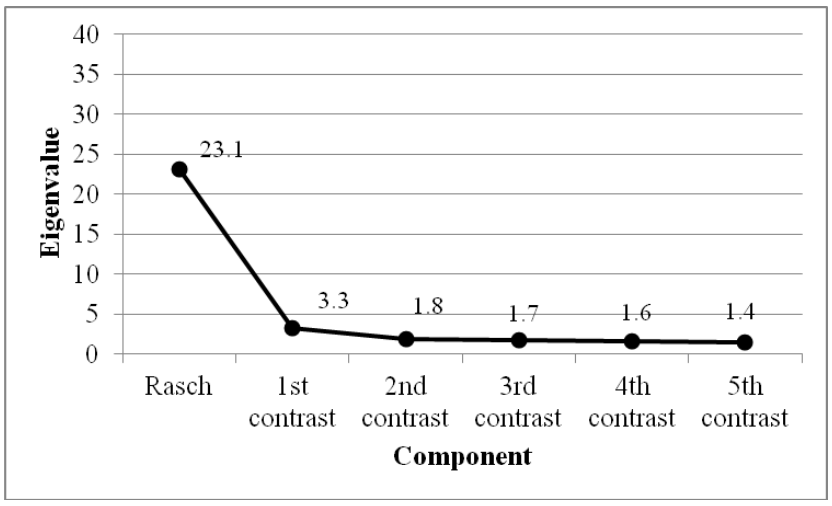


Figure 59. Scree plot for the use section

Table 66. Top 10 items with the largest positive and negative loadings (form section)

Positive loadings			Negative loadings		
Item	Loading	Difficulty	Item	Loading	Difficulty
mono-	0.36	-0.97	-ous	-0.22	0.33
un-	0.36	-1.65	-ation	-0.19	-0.49
in-	0.35	-1.37	-ling	-0.18	0.83
dis-	0.31	-1.59	-ency	-0.18	0.95
hyper-	0.31	-1.12	-et	-0.17	0.83
semi-	0.31	-1.18	-able	-0.17	-1.29
uni-	0.30	-1.46	-er	-0.15	-0.34
pro-	0.29	-1.58	-ant	-0.14	0.36
fore-	0.27	-0.78	-ess	-0.14	0.02
sub-	0.26	-2.00	-ify	-0.14	0.39

(Schumacker & Muchinsky, 1996; Spearman, 1904, 1910)). Together with the acceptable eigenvalue of the secondary dimension (2.6), it may be reasonable to conclude that the data from the form section is acceptably unidimensional.

Figure 58 presents the scree plot for the meaning section, showing that the first contrast had an acceptable eigenvalue (2.0), and the eigenvalues of other contrasts (1.8-1.5) reached an asymptote at the first contrast. This indicates that the data from the meaning section may be acceptably unidimensional.

Figure 59 presents the scree plot for the use section. This figure shows that the first contrast had an unacceptable eigenvalue (3.3), indicating that the data might have a secondary dimension. As with the form section, the contrast between strongly positively loading items and strongly negatively loading items on the first contrast was investigated. Table 67 presents the ten items with the largest positive and negative loadings on the secondary dimension (first contrast). This table shows that items with noun-making affixes had positive loadings on the first contrast, while items with adjective-making affixes had negative loadings on it, indicating that these two types of items might be measuring different constructs. However, it is practically useless to

Table 67. Top 10 items with the largest positive and negative loadings (use section)

Positive loadings			Negative loadings		
Item	Loading	Difficulty	Item	Loading	Difficulty
-ency(n)	0.44	0.55	-ory(adj)	-0.38	1.23
-age(n)	0.43	0.40	-some(adj)	-0.36	0.13
-th(n)	0.43	0.38	-esque(adj)	-0.35	0.71
-ness(n)	0.42	0.08	-ate(adj)	-0.32	0.02
-ation(n)	0.42	-0.32	-ly(adj)	-0.28	1.89
-cy(n)	0.40	0.24	-ary(adj)	-0.28	1.00
-ence(n)	0.40	-0.03	-atory(adj)	-0.26	1.51
-ion(n)	0.40	-0.18	-ar(adj)	-0.26	0.12
-ition(n)	0.35	-0.20	-y(adj)	-0.22	-0.58
-ure(n)	0.33	-0.08	-ent(adj)	-0.21	-0.79

Note: n=noun, adj=adjective.

create two subtests with one measuring noun-making affixes and the other measuring adjective-making affixes because the items would have the same answer on each subtest (e.g., on the subtest of noun-making affixes, the correct answer of all items would be *noun*). A close look at the ten strongly positively loading items showed that their infit and outfit mean-squares were all less than 1.0 (ranging between 0.55 and 0.87), indicating that they do not contradict the Rasch dimension but are rather predictive (e.g., *-ation*, *-ion*, and *-ition* may be redundant). On the other hand, nine of the ten strongly negatively loading items had the infit and outfit mean-squares of larger than 1.0 (ranging between 0.93 and 1.79), indicating that the data include an irregular pattern misfitting to the Rasch model. This may have increased the eigenvalue of the secondary component. As mentioned in Section 6.2.4.1, the Rasch fit analysis identified four of the ten items (*-ary, -ate, -ly,* and *-y*) as misfit, but the subsequent inspection indicated that they might be acceptable. Taken together, the eigenvalue of 3.3 for the first contrast of the use section may be negligibly larger than 3.0 which is considered to be an

acceptable threshold.

This section has investigated the structural aspect of construct validity by providing empirical evidence for unidimensionality of the data produced by the WPT items. Evidence from the three aspects (item correlations, fit analysis, and the PCA) largely indicates that the scores from the WPT are acceptably unidimensional.

6.2.4.4 Generalizability Aspect

The generalizability aspect of construct validity was investigated through item calibration invariance, person measure invariance, reliability, and invariance across administrative contexts (see Section 4.6.4 for details).

Item Calibration Invariance

The invariance of item calibrations was investigated by analysing uniform differential item functioning (DIF), an indication of unexpected behaviour by items showing that item calibrations vary across samples by more than the modelled error (Bond & Fox, 2007; Linacre, 2010a; Wolfe & Smith Jr., 2007). First, the DIF analysis was performed in order to examine whether the item calibrations from male (N=470) and female (N=580) test-takers[40] varied widely for each of the three sections. Welch's t-test revealed that statistically significant DIF was detected for eight items for the form section, five items for the meaning section, and seven items for the use section. Table 68 presents the Rasch difficulty estimates and Welch's t statistics for the items with significant DIF (p<.05) for the three sections.

[40] The data from 298 people who did not specify their gender were excluded from the analysis.

Table 68. DIF analysis for gender

Section	Item	Difficulty for males	Difficulty for females	Difference	*t*	*d.f.*	*p*
Form	-i	4.73	3.94	-0.79	-3.01	315	.003
	arch-	0.64	1.27	0.63	2.81	443	.005
	-ee	1.52	1.01	-0.52	-2.42	457	.016
	-ways	0.93	1.30	0.37	2.37	902	.018
	-ise	0.64	0.09	-0.56	-2.34	476	.020
	-ity	0.37	0.90	0.53	2.22	429	.027
	counter-	-0.65	-0.08	0.58	2.12	496	.035
	-ward	-0.11	0.39	0.50	2.01	472	.045
Meaning	counter-	1.14	1.70	0.56	2.94	719	.003
	-ling	2.22	1.80	-0.42	-2.44	739	.015
	-ster	0.45	-0.34	-0.79	-2.40	420	.017
	sub-	-0.22	0.54	0.76	2.31	401	.022
	-ian	-1.29	-2.50	-1.21	-2.04	466	.042
Use	-ate(adj)	0.42	-0.10	-0.52	-3.06	869	.002
	-ful(adj)	-0.20	-1.00	-0.80	-3.01	448	.003
	-y(adj)	-0.17	-0.63	-0.46	-2.64	931	.008
	-age(n)	0.24	0.68	0.44	2.50	786	.013
	-ible(adj)	0.15	-0.27	-0.42	-2.44	894	.015
	-able(adj)	0.42	-0.14	-0.55	-2.31	447	.021
	-ate(v)	-0.60	-0.20	0.40	2.28	906	.023

Note: adj=adjective, n=noun, v=verb.

The percentage of DIF items was above 5% (7.5% for the form section, 5.5% for the meaning section, and 12.5% for the use section) which may occur by chance given the nature of Type I error. This may have been due to a relatively large sample size (*N*=470 for males and *N*=580 for females), given that although 200 respondents per group may be necessary for adequate (>80%) power, a larger sample size than this may potentially identify a greater number of items as statistically significant DIF (Scott, et al., 2009; Tristan, 2006). A close look at the DIF items in Table 68 shows that the difference between the item difficulty estimates from males and females is around 0.5 logits and there is no systematic bias towards males or females, indicating that the DIF items have little effect on the accuracy of the measurement (Draba, 1977; Wright & Douglas, 1976).

DIF was also investigated for the participants' L1s. Given that at least 16

respondents are needed to obtain a person ability estimate of an accuracy within ± 1 logit with 95% confidence (Linacre, 1994), the following 15 L1 groups with 16 or more respondents were investigated: Arabic, Chinese, English, Filipino, French, German, Hindi, Indonesian, Japanese, Russian, Spanish, Tagalog, Tamil, Urdu, and Vietnamese. Pearson's product-moment correlation coefficients were calculated to examine the relationship between the item difficulty estimates from all the 1,348 respondents and those from each of the 15 L1 groups.[41] Table 69 shows that the item difficulty estimates from the respondents with 15 different L1s were significantly correlated to those from the overall respondents (average r = .842, ranging between .688 and .935). When corrected for attenuation, the average r was .954 (Min=.796, Max=1). This indicates that the overall item difficulties are highly generalisable across different L1s.

Table 69. Pearson's correlation coefficients between the item difficulty estimates from the overall participants and those from each of the 15 L1 groups

L1 group	*N*	Form		Meaning		Use	
Arabic	102	.921	(>1)	.916	(>1)	.874	(.982)
Chinese	46	.888	(.992)	.790	(.952)	.738	(.838)
English	226	.923	(>1)	.862	(>1)	.868	(.992)
Filipino	53	.910	(.995)	.852	(.979)	.860	(.966)
French	19	.820	(.944)	.814	(>1)	.752	(.875)
German	21	.847	(.981)	.865	(>1)	.788	(.917)
Hindi	93	.935	(>1)	.884	(>1)	.802	(.911)
Indonesian	60	.911	(>1)	.830	(.965)	.842	(.951)
Japanese	86	.893	(.971)	.794	(.883)	.749	(.823)
Russian	61	.900	(>1)	.847	(>1)	.802	(.911)
Spanish	47	.893	(.987)	.833	(>1)	.876	(.984)
Tagalog	28	.790	(.863)	.812	(>1)	.688	(.796)
Tamil	34	.811	(.896)	.886	(>1)	.736	(.827)
Urdu	67	.930	(>1)	.875	(.989)	.869	(>1)
Vietnamese	47	.787	(.865)	.837	(.990)	.801	(.890)

Note: Disattenuated correlation coefficients are presented in brackets. All the correlation coefficients are significant at p=.05.

[41] A *t*-test approach was not appropriate here, because there are 15 L1 groups to investigate, resulting in 105 ($_{15}C_2$) *t*-tests. Moreover, the majority of these L1 groups had far less than 200 respondents, a minimum number of respondents per group for adequate power (Scott, et al., 2009).

Finally, item calibration invariance was investigated for the test order in order to examine the effects of practice and fatigue. Although the form section always came first, the order of the meaning and the use sections was randomised. For the meaning and the use sections, the DIF analysis was performed to see whether significant differences were found between the item difficulty estimates from the respondents who worked on the meaning section prior to the use section (*N*=645) and those who worked on the two sections in the opposite order (*N*=703). Welch's *t*-test revealed that statistically significant DIF was detected for three items (4.1% of all items) for the meaning section and one item (1.8% of all items) for the use section. For both sections, DIF items accounted for less than 5% which may occur by chance. Table 70 presents the Rasch difficulty estimates and the Welch's *t* statistics for the items with significant DIF ($p < .05$) for the two sections, indicating that there is no systematic bias towards one group of respondents. Taken together, the item difficulty estimates for the meaning and the use sections are stable regardless of the test order, indicating that there is little effect of practice or fatigue on the accuracy of the measurement.

Table 70. DIF analysis for section order

Section	Item	Difficulty for meaning-first respondents	Difficulty for use-first respondents	Difference	*t*	*d.f.*	*p*
Meaning	fore-	0.25	-0.82	1.07	3.37	574	.001
	-ary	-1.33	-0.54	-0.79	-2.20	603	.028
	re-	-0.98	-1.86	0.87	2.05	608	.041
Use	-ant (adj)	0.45	-0.11	0.56	2.61	562	.010

Note: adj=adjective.

Person Measure Invariance

The invariance of person ability estimates was examined by analysing differential person functioning (DPF), an indication of unexpected behaviour by persons showing

that person measures vary across different situations by more than the modelled error (Bond & Fox, 2007; Linacre, 2010a). All items were divided into prefix and suffix items in order to examine whether person ability estimates were stable regardless of the two item groups (different ability estimates indicate that the two item groups are measuring different constructs). DPF was examined through a *t*-test approach. Table 71 presents the number of prefix and suffix items and the number of statistically significant DPF persons for each section.

Table 71. DPF analysis for prefixes vs. suffixes

	No. of prefix items	No. of suffix items	No. of DPF persons	%
Form	41	66	59	4.4
Meaning	39	34	46	3.4
Use	4	52	0	0.0

Note: % indicates the percentage of DPF persons in relation to the overall 1,348 respondents

Table 71 shows that the percentage of DPF persons was below 5% which may occur by chance. This indicates that person ability estimates are acceptably stable across affix types (prefixes and suffixes).

Reliability

A third way of investigating the generalizability aspect of construct validity was to examine the degree of reliability. Reliability was estimated by Rasch person and item reliabilities. Rasch person and item separations were also calculated because they are linear and range from zero to infinite. As missing data always decreases Rasch person reliability (Linacre, 2010a, p. 512), Cronbach's alpha was also calculated for an estimation of person reliability. Table 72 presents the reliability estimates for the three sections.

Table 72. Reliability estimates for the three sections

Section	Rasch person reliability	Rasch person separation	Rasch item reliability	Rasch item separation	Cronbach α
Form	.91	3.22	.99	10.18	.91
Meaning	.86	2.49	.99	8.29	.94
Use	.89	2.91	.99	9.09	.92

Table 72 shows high reliabilities for the three sections of the WPT, indicating a high degree of generalizability in terms of reproducibility of person ability and item difficulty estimates.

Invariance across Administrative Contexts

A final way of evaluating the generalizability aspect was to examine the stability of performance across administrative contexts. This was investigated by examining the relationships between the item difficulty estimates from the paper-based format in Study 1 (*N*=417) and those from the web-based format in Study 2 (*N*=1,348). Table 73 presents Pearson's correlation coefficients between them for the three sections.

Table 73. Correlation coefficients between item difficulty estimates from the paper-based and the web-based versions

Section	*r*	disattenuated *r*
Form	.868*	.954
Meaning	.684*	.782
Use	.615*	.680

*p<.05.

Considering that the paper-based format included a number of misfit items which were rewritten for Study 2, the correlation coefficients shown in Table 73 may be acceptably high. The use section obtained the lowest correlation perhaps due to different options between Studies 1 and 2: options were translated into Japanese in Study 1 (名詞, 動詞,

形容詞, and 副詞), whereas all options were written in English in Study 2 (noun, verb, adjective, and adverb). Adjectives and adverbs may have been difficult to differentiate for a number of participants in Study 2, because all the eight misfit items were either adjectives or adverbs (see Section 6.2.4.1). The participants in Study 1 may have been less confused with the differentiation between adjectives and adverbs, because 1) adjectives and adverbs are orthographically distinct from each other in Japanese, 2) many of the participants would be familiar with the parts of speech in Japanese instead of English, and 3) the eleven misfit items in the use section in Study 1 consisted of six adjective items and five noun and verb items. Taken together, the item difficulty estimates may be acceptably stable across administrative contexts.

This subsection has provided evidence for the generalizability aspect of construct validity of the WPT from four sub-aspects (item calibration invariance, person measure invariance, reliability, and invariance across administrative contexts). The empirical evidence largely indicates that the results from the present research are stable across different samples and situations.

6.2.4.5 External Aspect

The external aspect was investigated through correlations with external variables. For the evaluation of this aspect, the data from Study 1 were used, because 238 of the 417 participants who took the WPT also took the Japanese-bilingual version of the Vocabulary Size Test (VST) (Nation & Beglar, 2007) and 67 of them reported their TOEIC scores. It was hypothesised that the WPT scores would be positively correlated to the VST and TOEIC scores, but their correlations would be lower than the correlations between the scores from any two sections of the WPT. This is because the

three sections of the WPT measure different aspects of the same construct (receptive word part knowledge), whereas the VST and TOEIC measure different constructs which might partly be related to knowledge of word parts (see, for example, Mochizuki and Aizawa (2000) for discussion of the relationships between word part knowledge and vocabulary size, and Nagy, Berninger, and Abbott (2006) for discussion of the relationships between word part knowledge and reading ability). Table 74 presents a matrix of the Pearson's product-moment correlation coefficients between the WPT (N=417), VST (N=238), and TOEIC scores (N=67) without correction for attenuation due to unavailability of the reliability of the TOEIC scores.

Table 74. Correlation coefficients between the WPT, VST, and TOEIC scores

		WPT		
		Form	Meaning	Use
WPT	Form	-	.697*	.617*
	Meaning	.697*	-	.728*
	Use	.617*	.728*	-
VST		.449*	.467*	.375*
TOEIC		.426*	.428*	.327*

*p<.05.

Table 74 shows that the WPT scores positively correlated to the VST (r=.449, .467, .375) and the TOEIC scores (r=.426, .428, .327), but these correlations were lower than the within-WPT correlations (correlations between any two sections of the WPT (r=.697, .617, .728)). *Z*-tests were performed in order to determine whether the WPT-VST correlations (correlations between the WPT scores and the VST scores) and the WPT-TOEIC correlations (correlations between the WPT scores and the TOEIC scores) were significantly different. Tables 75 and 76 show that for all three sections, the within-WPT correlation coefficients were significantly higher than the WPT-VST and the WPT-TOEIC correlation coefficients (p<.05). The results support the above-

mentioned hypothesis (positive correlations for the WPT-VST and the WPT-TOEIC scores but lower correlations than the within-WPT correlations).

This subsection has looked at the relationships with external variables (VST and TOEIC scores). The results were supportive of the hypotheses about the relationships with these external variables.

Table 75. Difference between the within-WPT and the WPT-VST correlations

Section	within-WPT correlations (*N*=417)	WPT-VST correlations (*N*=238)	*z*	*p*
Form	r_{FM}=.697	r_{FV}=.449	4.63	.000
	r_{FU}=.617	r_{FV}=.449	2.90	.004
Meaning	r_{MF}=.697	r_{MV}=.467	4.35	.000
	r_{MU}=.728	r_{MV}=.467	5.12	.000
Use	r_{UF}=.617	r_{UV}=.375	3.99	.000
	r_{UM}=.728	r_{UV}=.375	6.49	.000

Note. F = form section, M = meaning section, U = use section, V = VST (e.g., r_{FM} = correlation coefficient between the scores of the form section and the meaning section).

Table 76. Difference between the within-WPT and the WPT-TOEFL correlations

Section	within-WPT correlations (*N*=417)	WPT-TOEIC correlations (*N*=67)	*z*	*p*
Form	r_{FM}=.697	r_{FT}=.426	3.03	.002
	r_{FU}=.617	r_{FT}=.426	1.97	.049
Meaning	r_{MF}=.697	r_{MT}=.428	3.01	.003
	r_{MU}=.728	r_{MT}=.428	3.48	.001
Use	r_{UF}=.617	r_{UT}=.327	2.83	.005
	r_{UM}=.728	r_{UT}=.327	4.36	.000

Note. F = form section, M = meaning section, U = use section, T = TOEIC.

6.2.4.6 Consequential Aspect

The consequential aspect of construct validity was investigated by examining whether the test results on which score interpretation and use are based might be affected by invalidity of the WPT. Invalidity may be caused by construct-irrelevant (measuring

something different from what it purports to measure) and construct under-representation difficulty (lacking in important construct-relevant items). The content relevance and representativeness of the WPT was supported by both theoretical argument and empirical evidence (see Section 6.2.4.1). This may be taken as supportive evidence for the consequential aspect because negative consequences are unlikely to be caused by the invalidity of the WPT.

Negative impact on consequences may also be caused by unfairness in test use (Messick, 1989, 1995, 1996). Unfairness was examined through item bias (different item difficulties across groups of respondents) in Rasch measurement (Smith Jr., 2004b; Smith, 1992). DIF analyses showed that the item difficulty estimates were stable regardless of gender, L1, and the section order (see Section 6.2.4.4), indicating that the WPT is unlikely to cause negative consequences due to unfairness (item bias). Unfairness may also occur in scoring when the responses are graded subjectively by judges; however, the WPT is written in a multiple-choice format which is free from variations in judge severity. This may be taken as supportive evidence for the consequential aspect of construct validity.

6.2.4.7 Responsiveness Aspect

The responsiveness aspect of construct validity was investigated through Rasch person strata. The person strata statistics for the three sections are presented in Table 77. This table shows that each section had a person strata index of greater than 2, which may be taken as supportive evidence for the responsive aspect of construct validity.

Table 77. Person strata for the three sections of the WPT

Section	Person strata
Form	4.63
Meaning	3.65
Use	4.33

Responsiveness may be decreased by a ceiling effect because able persons cannot demonstrate their gains from an experimental intervention such as teaching. As illustrated in Figures 51-53, the Rasch person-item maps showed that a number of participants obtained high person ability estimates, indicating a ceiling effect. The main reason for this is that the WPT was taken by a number of highly proficient learners including native English speakers. The ceiling effect was unlikely to be caused by under-representation of the construct because all the affixes that were measured in the WPT appeared in more than one word in the first 10,000 word families in the BNC word lists. Thus, persons with very high ability should be regarded as having sufficient knowledge of English word parts rather than showing evidence for invalidity due to a ceiling effect.

6.2.4.8 Interpretability Aspect

The interpretation of the scores may be facilitated by a Rasch person-item map for both norm-referenced and criterion-referenced assessment (see Section 4.6.8 for a detailed discussion). The simplest way of score interpretation may be to use raw scores, because teachers and learners have only to count the number of correct responses and do not need specialised computer software such as WINSTEPS. In order to investigate the adequacy of using raw scores for interpretation, Spearman's rank correlation coefficients between the raw scores and the Rasch person ability estimates were

examined (Table 78).[42] The raw scores were highly correlated to the Rasch person ability estimates (ρ > .9) regardless of the use of a missing data design. This indicates that the total number of correct responses may serve as a close approximation to the latent trait of word part knowledge. It should be noted here that the raw scores are only ordinal and are not on an interval scale. Thus, the difference between the score of 5 and 10 is not identical to the difference between the scores of 15 and 20.

Table 78. Correlation coefficients between the raw score and the Rasch person ability estimate for the three sections

Section	ρ
Form	.998*
Meaning	.997*
Use	.996*

*p<.05

Table 79 presents the relationships between the raw scores and the Rasch ability estimates for the three sections. The raw scores were converted to the percentage of correctly answered items that the participants had actually taken.[43] This table shows that, for example, a person who got 80% of the items correct for the form section has a Rasch ability of approximately 1.8 logits. This person has a greater probability of succeeding on any item with a Rasch difficulty estimate of less than 1.8 logits, and vice versa. In summary, a Rasch person-item map contributes to interpretability, and raw scores may also be used as a rough approximation of Rasch person ability estimates for convenience. This may be taken as supportive evidence for the interpretability aspect of construct validity of the WPT.

[42] Spearman's ρ was used because although the Rasch ability estimates were on an interval scale, raw scores were only ordinal.

[43] The Rasch ability estimate for each raw score category was based on the persons whose scores ranged between ±1% of the category. For example, the Rasch ability estimate for the raw score of 90% was calculated by averaging the person abilities of those who got 89-91% of the items correct.

Table 79. Conversion table of raw scores and Rasch ability estimates

Raw scores (%)	Form	Meaning	Use
100	6.3	6.0	5.3
90	2.8	3.0	2.6
80	1.8	1.9	1.7
70	1.2	1.2	1.1
60	0.6	0.6	0.5
50	0.0	0.0	0.0
40	-0.7	-0.8	-0.7
30	-1.6	-1.5	-1.3
20	-2.7	-2.5	-2.5

This section has investigated the validity of the WPT from the eight aspects of construct validity (content, substantive, structural, generalizability, external, consequential, responsiveness, and interpretability). The evidence provided in this section largely indicates that the WPT is a highly valid measure of receptive knowledge of English word parts. The subsequent section describes the procedure for developing new forms based on the item analysis in Study 2 and proposes a method for interpreting and reporting the results obtained from the new forms.

6.2.5 Creating New Forms

Two types of new test forms were created so that the WPT would be more useful to researchers, teachers, and learners. One involved two equivalent forms which had the same construct to be measured, the same test length, and the same distribution of item difficulties. The other involved three forms with different word parts at different difficulty levels (see Appendix H for all items of the revised WPT).

6.2.5.1 Equivalent Forms

The purpose of creating two equivalent forms is to measure learners' overall proficiency

in word part knowledge with an interval of time. Having two equivalent forms may be useful for research purposes because it allows a pre- and post-test design where the effects of teaching or learning tasks on word part knowledge may be investigated. Two equivalent forms were created by splitting the 118 word parts into halves so that the two forms 1) would have the same number of items for each of the three sections, 2) would not be statistically different in the mean and the variance of the item difficulties, and 3) would evenly include allomorphs (e.g., *-ation*, *-ion*, and *-ition*) because these items may be locally dependent (see Section 6.2.4.3). Table 80 presents the number of items in the three sections for the two forms. Although the two forms were different in number for the form and the meaning sections, the total number of items was the same.

Table 80. Number of items in the three sections for each form

Section	Form A	Form B
Form	54	53
Meaning	36	37
Use	28	28
Total	118	118

In order to statistically examine the homogeneity of variance of item difficulty between the two forms, Levene's tests were performed. The results showed that the null hypothesis of equal variances was not rejected for the three sections (F = 0.097, p = .756 for the form section; F = 0.058, p = .810 for the meaning section; and F = 0.243, p = .624 for the use section), indicating that the spread of item difficulties may be acceptably equal between the two forms. Subsequent *t*-tests (2-tailed) did not detect any significant differences in the mean item difficulties between the two forms for any of the three sections (Table 81). The effect sizes (r) were well below .20, indicating small differences between the two forms (Cohen, 1988, 1992). This may

indicate that the two forms are statistically equivalent.

Table 81. Comparison of the item difficulty between the two equivalent forms

	Form A		Form B		*t*	*d.f.*	*p*	*r*
	M	*SD*	*M*	*SD*				
Form	0.47	1.23	-0.47	1.12	0.417	105	.677	.041
Meaning	0.00	1.24	0.00	1.20	0.006	71	.995	.000
Use	-0.37	0.89	0.38	0.98	0.299	54	.766	.041

Another way of investigating the degree of equivalence of the two forms is to examine the item difficulty hierarchy for each form. Figures 60-62 illustrate the Rasch person-item maps of each form for the three sections. These figures show that the item difficulties are largely evenly distributed between Forms A and B for all three sections.

Finally, the quality of the two forms was investigated by examining their reliability estimates as calculated by the Spearman-Brown prediction formula (Brown, 1910; Spearman, 1910). Reliability was estimated based on the Rasch person reliability obtained in Study 2. Table 82 presents the estimated reliability and person strata of the two forms for the three sections. This table shows that the reliability estimates are .75 or above and the person strata estimates are larger than 2 which is the minimum requirement for a responsive test. It should be noted here that the estimated reliabilities are understated because of the missing data design in Study 2 (Linacre, 2010a). This indicates that the new forms may produce higher reliabilities than are estimated because they do not use a missing data design. Taken together, the two new forms may be taken as equivalent and reliable measures that are useful for research in the field of L2 vocabulary learning. (See Appendix I for the new forms of the WPT.)

The interpretation of the scores from the new test forms may be facilitated by Rasch measurement output including Rasch person ability estimates and Rasch person-

	<More able persons>			<More difficult items>					
			+	**Form A**			**Form B**		
	################ ################								
>2	###############		|	-i			-ster		
2	*############		+				-let		
	*###########		|	-atory	-ette		-esque		
	*###########	M	|	-ar	-eer		-ery		
	##########		|	-th	-fold				
	*##########		|	S -en	-or	-ways	-ite	-ory	-ure
1	*#########		+	-dom			-ance	-ancy	
	*########		|	-atic	-et	-ible	-ee	-most	-ese
	*#######		|	-an	-cy	-ency	arch-	-ion	-ary
	*######		|	-age	-ant	-ify	a-	-ian	-ling
	*####	S	|	-ence	-ent	-ize	-al	-hood	-ity
0	*###		|	M -ess -ous	-ic	neo-	mal- -ship	-ise	-ist
	*#		|	-ism	-ty	-ward	-ive	-some	-wise
	*#		|	-ate	-ation	circum-	-er		
	*##		|	counter-	be-	bi-	il-	-ish	sur-
	*#		|	en-			-ition	-y	em-
-1	*		+	co-	-ness	ir-	ab-	-ful	-less
	*#	T	|	S	-ment		fore-	anti-	-able
	*#		|	super-	mid-	post-	im-	mono-	trans-
	*#		|	dis-	non-	in-	hyper-	-ly	pre-
	*		|	de-	ex-	semi-	auto-	multi-	uni-
-2	*		+	inter-	mis-	un-	pro-	re-	sub-
<-2	*#		|	micro-					
	<Less able persons>		|	<Less difficult items>					

Note. # = 8 persons; * = 1 to 7 persons; M = mean of the person or item estimates; S = one standard deviation from the mean; T = two standard deviations from the mean.

Figure 60. Person-item map for the form section (Forms A and B)

	<More able persons>			<More difficult items>					
				Form A			Form B		
	*##########								
	##########								
>2	##########		|	-ette	in-		a-(not)	arch-	
2	*###		+	-et			-ling		
	*##		|	de-					
	*###		|						
	*#		|	counter-	neo-		-ship		
	*###		|				mal-	pro-	
	*##		| S				-ite	sur-	uni-
1	*#	S	+	ex-	-fold		ab-	-let	
	*##		|	inter-					
	*		|	-dom	-ism	super-			
	*#		|	circum-					
	*		|				-less	sub-	
	*		|	co-	non-			hyper-	mono-
0	*		+ M	bi-	-i	-ible	a-(toward)	-ful	-hood
				post-			il-	im-	
	*		|	dis-	ir-	un-	fore-	-ster	-wise
	*		|	-ward			auto-		
	*		|	-ent			-most		
	*	T	|				anti-		
	*		|	-ant	micro-	semi-	-able	-ary	
-1	*		+	-or			multi-	trans-	
	*		| S	-en	mis-		pre-		
	*		|	-ess			-er	re-	
	*		|	-ways			-ee		
	*		|	-an					
	*		|	-eer	mid-				
-2	*		+				-ian	-ist	
<-2	*#		|	-th			-ese		
	<Less able persons>			<Less difficult items>					

Note. # = 15 persons; * = 1 to 14 persons; M = mean of the person or item estimates; S = one standard deviation from the mean; T = two standard deviations from the mean.

Figure 61. Person-item map for the meaning section (Forms A and B)

	<More able persons>			<More difficulty items>			
				Form A		Form B	
	*###########						
	###########						
>2	###########						
2	######		+	T		-ly(adj)	
	*#######			-atory(adj)			
	*#######						
	*#########			be-(v)		-ory(adj)	
	*########						
	*########					a-(adv)	-ary(adj)
1	#########	M	+	S	-ty(n)		em-(v)	
	#########			-ency(n)		-esque(adj)	-y(n)
	*######			-en(v)	en-(v)	-ise(v)	
	########			-age(n)	-th(n)	-ancy(n)	
	#######			-ize(v)	-ify(v)	-al(n)	-ity(n)
	*#########			-ant(adj)	-cy(n)	-ery(n)	-some(adj)
				-ar(adj)	-ate(adj)	-able(adj)	-ion(n)
0	*#######		+	M	-ence(n)	-ness(n)	-less(adj)	-ure(n)
	*##########			-ible(adj)	-ic(adj)	-ance(n)	-ition(n)
	*#########			-atic(adj)	-en(adj)	-al(adj)	-ive(adj)
	*########			-ate(v)	-ation(n)	-ly(adv)	-y(adj)
	*#########	S		-ment(n)		-ful(adj)	
	*########			-ent(adj)			
-1	*######			S	-ous(adj)		-ish(adj)	
	*#####						
	*##						
	*###					-ary(n)	
	###			-ant(n)			
	*#						
-2	*#		+	T	-ent(n)		-ee(n)	
<-2	*#####			-or(n)		-er(n)	
	<Less able persons>			<Less difficult items>			

Note. # = 5 persons; * = 1 to 4 persons; M = mean of the person or item estimates; S = one standard deviation from the mean; T = two standard deviations from the mean; n = noun; v = verb; adj = adjective; adv = adverb.

Figure 62. Person-item map for the use section (Forms A and B)

Table 82. Estimated reliability and person strata of the new forms

Section	Target No. of items	Estimated reliability	Estimated person strata
Form	53	.83	3.32
Meaning	36	.75	2.65
Use	28	.80	3.02

item maps. The Rasch measurement has the advantage of having the scores on an interval scale and thus allowing the comparison between multiple groups of learners and the investigation of the development of word part knowledge through statistical tests such as a *t*-test and an ANOVA. Raw scores, ordinal as they are, may be used for convenience by referring to the conversion table presented earlier (Table 79).

The equivalent forms may be useful for research purposes, but have the disadvantage of measuring learners' knowledge of difficult items that are unlikely to be known by many learners (e.g., *-ette*). The subsequent subsection describes the procedure for creating another type of test form that may be more useful for teachers and learners.

6.2.5.2 Forms with Different Difficulty Levels

This section describes the procedure for creating the Word Part Levels Test (WPLT) which has three different forms with different difficulty levels. The primary purpose of creating these forms is to offer a diagnostic word part test to determine the level of the learners and raise awareness of word parts to learn.

The WPLT was created by classifying the 118 word parts into three difficulty levels by averaging the item difficulty from each section.[44] Level 1 consists of the 40

[44] For example, the difficulty of the word part *-able* (-0.68) was obtained by averaging the difficulties from the form (-1.29), the meaning (-0.83), and the use (0.08) sections. The difficulty of *anti-* (-0.98) was derived by averaging the difficulties from the form (-1.29) and the meaning (-0.66) sections because this item was not measured in the use section.

least difficult word parts, Level 3 consists of the 39 most difficult ones, and Level 2 consists of 39 word parts of middle difficulty. Table 83 presents a summary of the three forms.

Table 83. Number of word parts and items in the three forms

Form	Word part level	No. of word parts	No. of items			
			Form section	Meaning section	Use section	Total
A	Easy	40	40	34	13	87
B	Middle	39	37	21	21	79
C	Hard	39	38	18	22	78

Each section of the three forms had a sufficient number of items to achieve estimated person strata of greater than 2 (16 items for the form section, 18 for the meaning section, and 10 for the use section) based on the Spearman-Brown prediction formula (Brown, 1910; Spearman, 1910). The WPLT had three levels instead of four or more levels in order to make sure that the estimated person strata of every section in every form would exceed 2 which means that the test may statistically distinguish at least two person levels. Table 84 presents the average item difficulty for the three forms. This table shows that Form A had easy items, Form C had difficult items, and Form B had items of middle difficulty.

Table 84. Average item difficulty for the three forms

Form	Form section	Meaning section	Use section
A	-0.77	-0.74	-1.15
B	-0.12	0.03	0.13
C	0.96	1.37	0.55

Figures 63-65 illustrate the Rasch person-item maps of each form for the three sections. These figures also show that an easier form tends to have easier items for all three sections, and vice versa.

	<More able persons>				<More difficult items>					
			+		Form A		Form B		Form C	
	##########									
	##########									
>2	##########		|						-i	-ster
2	*##########		+							-esque	-let
	*##########		|						-atory	-ette
	*##########	M	|				-eer		-ar	-ery
	##########		|				-th		-fold	-ory
										-th	-ure
	*##########		|	S	-or		-en	-ways	-en	-ite
1	*##########		+							-ance	-ancy
	*##########		|		-ee	-ese	-ible	-most	arch-	-dom
						-an	-ary	-atic		-ary	-cy
	*##########		|						-ency	-et
										-ion	
	*##########		|		-ian	-ant	-ant		a-	-age
										-ify	-ling
	*##########	S	|			-ent	-al	-ence	-ity	-ize
								-ent	-hood		
						-ess	-ist	-ous		-ic	-ise
0	*##########		|	M					mal-	neo-
										-ship	
								-ism	-ive		
	*##########		|				-some	-ty		
								-ward	-wise		
	*##########		|		-er		-ate	-ation	counter-	
								circum-			
	*##########		|		-ish		bi-	il-	be-	sur-
	*##########		|			-y	em-	en-		
								-ition	-y		
-1	*		+			co-	-ful	ab-	-ful		
						ir-	-ness	-less			
	*##########	T	|	S		-able	anti-				
						fore-	-ment				
	*##########		|		mid-	mono-	super-			
						post-	trans-				
						dis-	hyper-	-ly		in-	
	*##########		|		im-	-ly				
						non-	pre-				
	*		|		auto-	multi-	de-	ex-		
						semi-		uni-			
						inter-	mis-	pro-			
-2	*		+			re-	sub-				
						un-					
<-2	*##########		|		micro-					
	<Less able persons>		|		<Less difficult items>					

Note. # = 8 persons; * = 1 to 7 persons; M = mean of the person or item estimates; S = one standard deviation from the mean; T = two standard deviations from the mean.

Figure 63. Person-item map for the form section (Forms A, B, and C)

	<More able persons>			<More difficult items>					
				Form A		Form B		Form C	
	*##########								
	##########							a-(not)	arch-
>2	##########		|					-ette	in-
2	*###		+					-et	-ling
	*##		|			de-		
	*###		|					
	*#		|					counter-	neo-
								-ship	
	*###		|			pro-		mal-
	*##		|	S		uni-		-ite	sur-
1	*#	S	+			ab-	ex-	-fold	-let
	*##		|		inter-			
	*		|			-ism	super-	-dom	
	*#		|			circum-		
	*		|		sub-	-less		
	*		|		co-	hyper-		
						mono-	non-		
						im-	post-	bi-	il-	a-(toward)	-i
0	*		+	M			-ful	-hood		
							-ible		
	*		|		dis-	fore-	wise		-ster	
						ir-	un-				
	*		|		auto-		-ward			
	*		|		-ent			-most		
	*	T	|		anti-					
	*		|		-able	-ant			-ary	
						micro-	semi-				
-1	*		+			multi-	-or				
						trans-					
	*		|	S	mis-	pre-	-en			
	*		|		-er	-ess				
						re-					
	*		|		-ee		-ways			
	*		|		-an					
	*		|		mid-		-eer			
-2	*		+			-ian	-ist				
<-2	*#		|		-ese		-th			
	<Less able persons>			<Less difficult items>					

Note. # = 15 persons; * = 1 to 14 persons; M = mean of the person or item estimates; S = one standard deviation from the mean; T = two standard deviations from the mean.

Figure 64. Person-item map for the meaning section (Forms A, B, and C)

<More able persons>				<More difficulty items>		
				Form A	**Form B**	**Form C**
	*##########					
	##########					
>2	##########	|				
2	######	+	T		-ly(adj)	
	*#######	|				-atory(adj)
	*#######	|				
	*#########	|				-ory(adj)
	*########	|				be-(v)
							a-(adv)
	*########	|				-ary(adj)
1	########	M	+	S			
	#########	|			y(n) em-	-esque(adj)
						(v)	-ency(n)
	*######	|				-ty(n)	-en(v)
						en-(v)	-ise(v)
							-ancy(n)
	########	|					-age(n)
							-th(n)
							-ity(n)
	#######	|					-ize(v)
							-ify(v)
	*#########	|			-ant(adj)	-cy(n)
						-al(n)	-ery(n)
					-ness(n)	-some(adj)	-ar(adj)
0	*#######	+	M		-able(adj)	-less(adj)	-ure(n)
						-ate(adj)	-ion(n)
						-ence(n)	
						-ible(adj)	-ic(adj)
	*##########	|				-ition(n)	-ance(n)
	*#########	|				-en(adj)	
							-atic(adj)	
							-ation(n)	
	*########	|				-ate(v)	
							-al(adj)	
							-ive(adj)	
					-ly(adv)	-ent(adj)	
	*#########	S	|		-y(adj)		
					-ment(n)		
					-ful(adj)		
	*########	|					
-1	*######	S	S		-ish(adj)	-ous(adj)	
	*#####	|					
	*##	|					
	*###	|			-ary(n)		
	###	|			-ant(n)		
	*#	|					
-2	*#	+	T		-ent(n)		
					-ee(n)		
	*#####	|			-er(n)		
<-2					-or(n)		
	<Less able persons>			<Less difficult items>		

Note. # = 5 persons; * = 1 to 4 persons; M = mean of the person or item estimates; S = one standard deviation from the mean; T = two standard deviations from the mean; n = noun; v = verb; adj = adjective; adv = adverb.

Figure 65. Person-item map for the use section (Forms A, B, and C)

The usefulness of the word part classification (Easy, Middle, and Hard) was investigated by examining whether less difficult word parts would be worth learning earlier than more difficult ones. In so doing, the frequency of the word parts was investigated, because more frequent word parts appear in a greater number of words and thus may contribute to learning many words that include the word parts. Table 85 presents the average lemma and token frequency of the word parts at each level.

Table 85. Average word part frequency for each level

Level	Ave. lemma frequency	Ave. token frequency
Easy	295	101,478
Middle	96	77,586
Hard	65	61,264

Table 85 shows that for both types of frequency counts, an easier level tends to have more frequent word parts, indicating that learning word parts from easy to hard levels may be effective.

The quality of the word parts at each level was also investigated by examining the relationship with Bauer and Nation's (1993) seven levels of affixes. It was hypothesised that an easier form would contain a greater number of word parts at a lower level than a harder form, and vice versa. Figure 66 presents the number of word parts in each form according to Bauer and Nation's affix level.[45] This figure shows a weak tendency that an easier form contains a greater number of word parts at lower levels of Bauer and Nation's affix levels, and vice versa. (See Appendix J for the WPLT.)

[45] Levels 1 and 2 are omitted here because they deal with word types and inflectional suffixes which are not the focus of the present research. Level 7 was also omitted because no information was provided on frequency, productivity, predictability, or regularity.

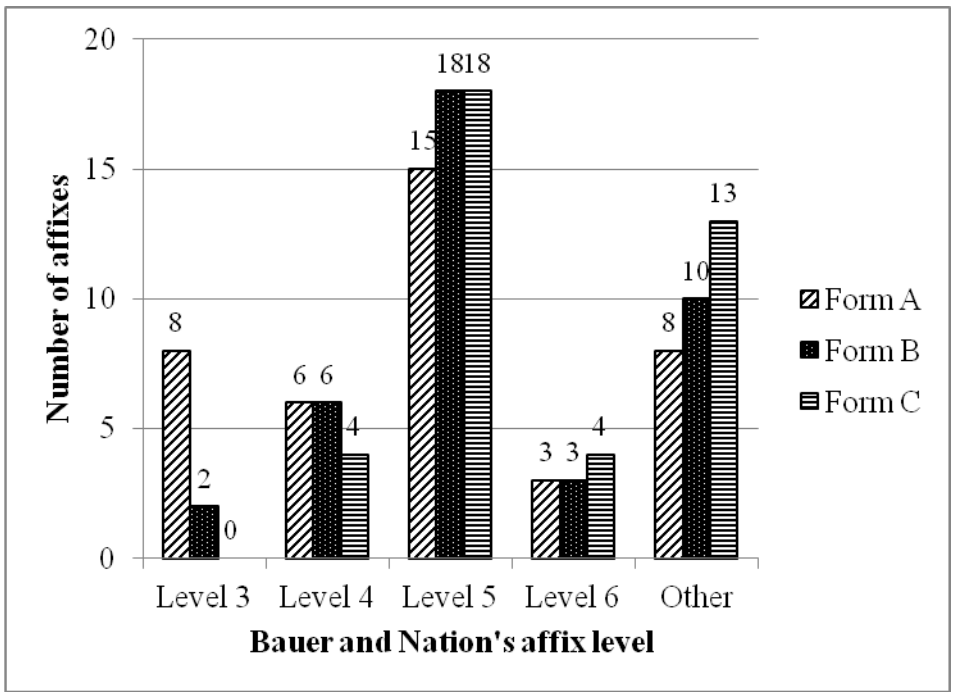


Figure 66. Bauer and Nation's affix levels and three new forms

The scores may be interpreted based on the percentage of correct answers instead of Rasch measurement, because the purpose is to diagnose how many word parts at each level a learner knows, and not to estimate his or her ability in word part knowledge. The use of percentage will provide learners with diagnostic information about how many word parts they need to know to move on to the next level.

For practical use of the WPLT, diagnostic feedback needs to be easy for learners and teachers to understand so that learners' weaknesses in word part knowledge may be clearly indicated. To meet this need, a bar graph may be useful because the information is visually presented and intuitively interpretable. Suppose Learner A took Form A and got 90% of the items correct for the form and the meaning sections and 60% correct for the use section. This learner's scores may effectively be reported in the bar graph as shown in Figure 67. The horizontal axis indicates the WPLT sections, and the vertical

axis indicates the percentage of correct answers. The bar graph shows that this learner demonstrated good knowledge of word part forms and meanings but his or her use knowledge is not sufficient. Thus, this learner may need to focus on the learning of word part use to move on to the next level.

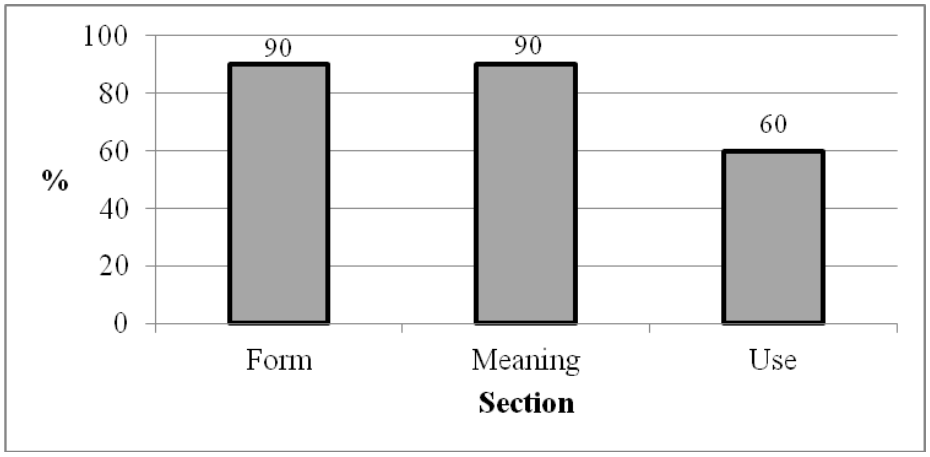


Figure 67. Score report (Learner A)

This section has discussed the procedure for creating new forms of the WPT and the WPLT and ways in which the scores may be interpreted and reported to learners. Two equivalent forms may be useful in investigating learners' overall proficiency in word part knowledge. Three forms with different difficulty levels may be effective in finding out where learners are experiencing difficulty and provide them with diagnostic feedback on how they can improve their knowledge of word parts.

6.3 Discussion

One of the important features of the WPT is its comprehensiveness (measuring multiple aspects of word part knowledge). The WPT measured three aspects of receptive word part knowledge which were included in previous studies that discussed what is involved

in affix knowledge (Bauer & Nation, 1993; Nation, 2001; Tyler & Nagy, 1989). This is in line with recent L2 vocabulary studies which measured multiple aspects of vocabulary knowledge for investigating 1) the relationships between the different aspects of vocabulary knowledge (Schmitt, 1998, 1999; Schmitt & Meara, 1997), 2) the relationships between vocabulary knowledge and other skills such as reading ability (Qian, 1999), and 3) the effects of learning tasks (Webb, 2005, 2007a, 2007b, 2007c, 2009). The WPT may also be a useful tool to shed light on how these three issues relate to knowledge of word parts.

For example, first, it is useful to examine the correlation coefficients among the three aspects of word part knowledge. Table 86 presents a matrix of Pearson's product-moment correlation coefficients among the Rasch person ability estimates from Study 2.

Table 86. Correlation coefficients between the WPT scores

	Form	Meaning
Meaning	.697* (.788)	-
Use	.617* (.686)	.728* (.832)

*p<.05.
Note: Adapted from Table 78. Disattenuated correlation coefficients are in brackets.

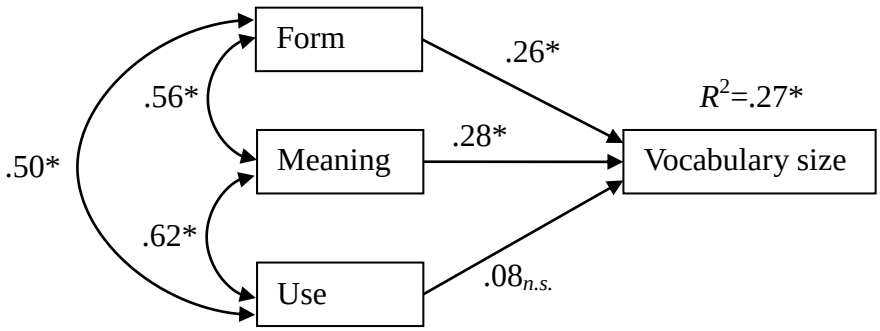
Table 86 shows that the three aspects of word part knowledge are positively and moderately correlated to each other. This indicates that these three aspects may develop at a similar pace, perhaps because learners gain knowledge of affixes by relating the word part form to its meaning and grammatical function. This result is in line with the findings of Webb (2009) indicating that learners gain different aspects of vocabulary knowledge including orthography, meaning, and grammatical function as a result of word-pair leaning.

Second, the relationships between the three aspects of word part knowledge and

vocabulary size were investigated using the data from 238 participants in Study 1 who took both the WPT and the Japanese bilingual version of the VST. Pearson's product-moment correlation coefficients between the Rasch person ability estimates from the WPT and the VST were .449 for the form section, .467 for the meaning section, and .375 for the use section (see Table 74). This supports the findings of Schmitt and Meara (1997) and Mochizuki and Aizawa (2000) demonstrating that affix knowledge correlated with vocabulary size (Pearson's *r* ranged between .35 and .65). This positive correlation may be due to the interaction between word part knowledge and vocabulary size where larger vocabulary size provides learners with a greater chance of recognising word parts which in turn facilitates the learning of new words that include the word parts they already know.

In order to further examine the relationships between word part knowledge and vocabulary size, a multiple regression analysis was performed with the dependent variable being the Rasch person ability estimates from the VST and the independent variables being the Rasch person ability estimates from the three sections of the WPT. A path diagram of the results is presented in Figure 68 (without correction for attenuation).[46] This figure shows that knowledge of word part forms (β = .26) and meanings (β = .28) significantly contributes to vocabulary size, whereas use knowledge does not. This may be because the VST measures the form-meaning relationships of words and does not focus on grammatical function. A combination of the three aspects of word part knowledge accounted for about a quarter of the variability of vocabulary size (R^2=.27). Given that vocabulary size is affected by many other factors such as the skill of guessing from context and general language proficiency, this coefficient of

[46] No serious sign of multi-collinearity was detected. The variance inflation factor (VIF) was 1.53 for the form section, 1.87 for the meaning section, and 1.71 for the use section, which are well below 10 which is generally taken as the threshold for multi-collinearity.



*p<.05. *n.s.*=not significant

Figure 68. Relationships between the three aspects of word part knowledge and vocabulary size

determination may be considered high. Another point to be made with Figure 68 is that knowledge of affix forms had a significant contribution to vocabulary size. This indicates the importance of measuring the form aspect which was not examined in previous studies (Mochizuki & Aizawa, 2000; Schmitt & Meara, 1997). Taken together, the results showed that word part knowledge was related to vocabulary size, but the relationship varied according to different aspects of word part knowledge.

Finally, future research may investigate the effects of teaching and learning tasks on different aspects of word part knowledge. For example, learning word parts with repetition and context may be effective, but may have different effects on word part knowledge, given that the quality of context has a greater effect on gaining knowledge of meaning while the number of encounters has a greater effect on gaining knowledge of form (Webb, 2007a, 2007c, 2008).

6.4 Summary

This chapter aimed to validate the WPT so that it would be widely available to

researchers, teachers, and learners. In so doing, two main studies were conducted. Study 1 aimed to identify poorly written items of the WPT with 417 Japanese learners of English with a wide range of proficiency levels. Six different forms were created in a paper-based format and were randomly distributed to the participants. Rasch analysis identified eleven misfit items for the form section, nine for the meaning section, and eleven for the use section. These misfit items were inspected and rewritten where necessary. Study 2 examined the validity of the revised WPT from the eight aspects of construct validity (content, substantive, structural, generalizability, external, consequential, responsiveness, and interpretability) with 1,348 participants with different L1 backgrounds. Table 87 summarises the evidence provided for the validity argument. On the whole, both the logical argumentation and the empirical evidence from the eight aspects indicated a high degree of validity of the WPT. It should be noted that five items in Table 88 were unacceptable in terms of Rasch fit analysis and need watching for future use of the test.

For future use of the WPT, two equivalent forms were created by splitting the items into halves so that each form had a total of 118 items with the same spread of difficulty. These new forms are useful for researchers because the effects of teaching and learning tasks may be effectively investigated by a pre- and post-test design. In order for the WPT to be more useful to teachers and learners, the WPLT was created by classifying the 118 word parts into three difficulty levels. The WPLT may determine whether the learner has mastered word parts at a particular level.

The scores obtained from the WPT are highly interpretable in the context of Rasch measurement. For more convenient interpretations, conversion tables (see Table 79) between raw scores and Rasch person ability estimates are provided. The scores may be

effectively reported to learners using a bar graph which presents learners' weaknesses visually. Taken together, it should be reasonable to conclude that the WPT is a highly valid measure for assessing word part knowledge and useful for both research and practical purposes.

Table 87. Summary of evidence provided for the WPT

Aspects	Sub-category	Evidence provided
Content	1. Content relevance	Test specifications
	2. Representativeness	Rasch item strata
		Rasch person-item map
	3. Technical quality	Rasch item fit analysis
Substantive		Test of difficulty hypotheses
		Rasch person fit analysis
Structural		Dimensionality analysis
Generalizability	1. Item calibration invariance	DIF analysis for gender, L1, and test order
	2. Person measure invariance	DPF analysis for affix types (prefixes vs. suffixes)
	3. Reliability	Rasch person separation and reliability
		Rasch item separation and reliability
	4. Invariance across administrative contexts	Comparison between person ability estimates from paper- and web-based format
External		Correlation with vocabulary size (as measured by the VST) and general language proficiency (as measured by TOEIC)
Consequential		Analysis of sources of invalidity
		Item bias
Responsiveness		Person-item map (ceiling effects)
		Person strata
Interpretability		Person-item map
		Conversion of raw scores and Rasch person ability estimates

Table 88. Misfit items in Study 2

Section	Item
Form	-th
	-i
Meaning	-ling
Use	-ly (adjective)
	a- (adverb)

CHAPTER 7

GENERAL DISCUSSION AND CONCLUSION

This chapter reviews the research presented in this thesis, and discusses pedagogical implications and directions for future research.

7.1 Review of the Research

The purpose of the present thesis was to create diagnostic tests of English vocabulary learning proficiency (tests measuring how proficiently words are learned). Among the various types of knowledge and strategies that may facilitate vocabulary learning, this thesis focused on the skill of guessing from context and knowledge of word parts because they are teachable and are used most frequently when learners deal with unknown words (Baumann et al., 2002; de Bot, et al., 1997; Mori, 2002; Mori & Nagy, 1999; Paribakht & Wesche, 1999). In this thesis, a successful attempt was made to create and validate a guessing from context test (GCT) and a word part test (WPT). The most important feature of these two tests was that each of the tests measured multiple aspects of the construct in order to provide learners with diagnostic information on their weaknesses in vocabulary learning. Another important feature was that the tests were easy to complete and grade because they were written in a multiple-choice format. This would allow learners to receive prompt feedback on their weaknesses.

The GCT was designed to measure knowledge of three aspects of guessing from context (identifying the part of speech of the unknown word, identifying the contextual clue, and deriving meaning) based on previous studies of the strategies for guessing

from context (Bruton & Samuda, 1981; Clarke & Nation, 1980; Williams, 1985). It consisted of three sections (part of speech, contextual clue, and meaning) measuring each of the three aspects. The test words to be guessed were randomly selected from low-frequency words (words listed between the 11th and 14th 1,000 word families in the BNC word lists) and were replaced by nonsense words. Each test word was embedded in a context that consisted of between 50 and 60 high-frequency words (words that were listed in the first 1,000 word families in the BNC word lists or words that were likely to be known to learners at a beginner level based on a series of pilot studies). The three aspects of guessing from context were measured separately in three sections. The first section was to choose the correct part of speech of the test word from a set of four options (noun, verb, adjective, and adverb). The second section was to choose the word or phrase that was most helpful in determining the meaning of the test word from three options. The last section was to choose the closest meaning of the test word.

A total of 60 items were created for the GCT and the quality of these items was examined through data from 428 Japanese learners of English with a wide range of proficiency levels. Rasch analysis revealed that eleven of the 60 items would be unacceptable and these eleven items were excluded from the GCT. The validity of the GCT with the 49 acceptable items was evaluated from eight aspects of construct validity (content, substantial, structural, generalizability, external, consequential, responsiveness, and interpretability) (Medical Outcomes Trust Scientific Advisory Committee, 1995; Messick, 1989, 1995). The results indicated that (1) the items were relevant to and representative of the construct being measured and showed good fit to the Rasch model; (2) the item difficulties and person abilities were generally consistent with *a priori*

hypotheses; (3) the items showed a high degree of unidimensionality; (4) item difficulty and person ability estimates were acceptably stable in terms of gender, L1, test order, internal consistency, and test length; (5) the scores from the GCT significantly correlated with those from a productive version of the GCT where the test-takers wrote answers instead of choosing answers from a set of options (ρ=.77-.91) and self-reported TOEIC scores (r=.24-.46); (6) item bias (different item difficulties across groups of test-takers) was not observed; (7) the GCT was able to distinguish two statistically different levels (the average Rasch person strata of larger than 2); and (8) the scores were highly interpretable with Rasch person-item maps and conversion tables between raw scores and Rasch person ability estimates. Taken as a whole, the GCT is a highly valid and reliable measure of the skill of guessing from context. The results also indicated that 20 items would be needed for achieving the minimum person strata estimate of 2. Two new equivalent forms each with 20 items were created in order to allow a pre- and post-test design where researchers and teachers can investigate learners' development of the skill of guessing from context. The GCT is also of great practical use because it diagnoses where learners find difficulty in guessing from context. Some may improve their guessing skill by gaining knowledge of part of speech, while others may do so by learning various types of contextual clues. Diagnostic feedback may be provided effectively using a bar graph where learners can visually recognise their weaknesses.

The WPT was designed to measure knowledge of three aspects of word part knowledge (recognising written forms of word parts, recognising meanings of word parts, and recognising syntactic properties that word parts have) based on previous studies of what is involved in knowing affixes (Bauer & Nation, 1993; Nation, 2001; Tyler & Nagy, 1989). A word part was defined as a bound morph that attaches to a free

morph. The WPT measured 118 word parts that appeared in more than one word in the first 10,000 word families in the BNC word lists. It consisted of three sections (form, meaning, and use) measuring each of the three aspects of word part knowledge. For the form section, test-takers must choose the correct word part form from four options with three distractors which were real but meaningless sequences of letters in English (e.g., *-ique* in *technique*). The meaning section required test-takers to choose the correct meaning of the word parts from four options. For the use section, test-takers must choose the part of speech that the word part makes from four options (noun, verb, adjective, and adverb). The meaning and the use sections presented two example words for each item in case one is unknown. The example words were the most frequent, semantically transparent, and regularly connected words to maximise the likelihood that they would be known to test-takers.

The quality of the WPT was examined in two studies. Study 1 was conducted with 417 Japanese learners of English with a wide range of proficiency levels in order to identify poorly written items. Rasch analysis detected eleven misfit items for the form section, nine for the meaning section, and eleven for the use section. These misfit items were inspected and rewritten where necessary. Study 2 evaluated the validity of the revised WPT from the eight aspects of construct validity (content, substantive, structural, generalizability, external, consequential, responsiveness, and interpretability) with 1,348 participants with different L1 backgrounds. The results indicated that (1) the items were relevant to and representative of the construct being measured and showed good fit to the Rasch model; (2) the item difficulties and person abilities were generally consistent with *a priori* hypotheses; (3) the items were acceptably unidimensional; (4) item difficulty and person ability estimates were acceptably stable in terms of gender, L1, test

order, affix types (prefix vs. suffix), internal consistency, and test format (paper- vs. web-based); (5) the scores from the WPT significantly correlated with those from the VST (*r*=.38-.47) and self-reported TOEIC scores (*r*=.33-.43); (6) item bias was not observed; (7) the WPT was able to distinguish two statistically different levels; and (8) the scores were highly interpretable with Rasch person-item maps and conversion tables between raw scores and Rasch person ability estimates. Taken as a whole, the WPT is a highly valid and reliable measure of word part knowledge. Two new equivalent forms of the WPT were created by splitting the 118 word parts into halves in order to allow a pre- and post-test design. For more practical use of the test, the Word Part Levels Test (WPLT) was created by classifying the 118 word parts into three different levels of difficulty. This allows teachers to quickly examine whether their students need to work on easy or difficult word parts and which aspects of word part knowledge need to be learned. As with the GCT, diagnostic feedback may be provided effectively by using a bar graph where learners can visually recognise their weaknesses.

7.2 Limitations

The GCT and the WPT are not perfect tools for measuring the skill of guessing from context and knowledge of word parts. The GCT measures knowledge of three aspects of guessing from context (grammatical knowledge, discourse knowledge, and deriving meaning). However, as previous studies (de Bot, et al., 1997; Haastrup, 1985, 1991; Nassaji, 2003) argue, other types of knowledge such as world knowledge and L1 knowledge may affect the success in guessing. These types of knowledge are not measured in the GCT because they are not teachable and available in every context, but

measuring these types of knowledge may be necessary for a more comprehensive test of guessing from context.

The WPT measures three aspects of word part knowledge (recognition of word part forms, knowledge of word part meaning, and knowledge of word part use). However, knowledge of word parts may involve other aspects such as knowing the changes of word forms that occur when an affix is attached (e.g., attaching the affix *-ness* to the word *happy* causes a change in spelling and the word *happiness* results) and knowing which word classes certain affixes can take (e.g., *repeatise* is impossible because the affix *-ise* cannot attach to a verb) (Bauer & Nation, 1993; Nation, 2001). These aspects of word part knowledge are not measured in the WPT because the present research focused on receptive rather than productive knowledge of word parts. However, measuring productive knowledge may also be important because learners may have limited ability to produce appropriate derivatives (Schmitt & Zimmerman, 2002).

This thesis has provided initial evidence for the validity of the GCT and the WPT. Although the results generally indicate that these two tests are valid and reliable, further research is still needed for investigating the validity of the tests. The GCT was validated with Japanese learners of English and future research may evaluate its validity with learners with other L1 backgrounds. It was also validated using a paper-based format and future research may use a computer-based format which may effectively counterbalance practice and fatigue effects by randomising the order of the items and control learners' test-taking strategies such as going back to previous sections to change their answers.

The WPT was validated in two studies with learners with a wide variety of L1 backgrounds and with both paper- and computer-based formats. Based on Study 1 poor

items were rewritten, but Study 2 indicated that five items were unacceptable in terms of Rasch fit analysis. Research is needed to further examine the quality of these items.

7.3 Suggestions for Future Research

The GCT and the WPT may make a number of future studies possible. First, the effects of teaching may be investigated with the GCT and the WPT. Previous studies on guessing from context (Fukkink & de Glopper, 1998; Kuhn & Stahl, 1998; Walters, 2006) generally indicate that teaching may result in improvement of the skill of guessing, but are not consistent with the relative efficacy of teaching methods. One of the reasons for this inconsistency may be that only one aspect of guessing (deriving the meaning of an unknown word) was measured. The GCT measures three aspects of guessing, and thus can examine what teaching methods contribute to improving each aspect of guessing. Teaching a variety of contextual clues might be effective because learners may increase discourse knowledge which contributes to improvement of guessing skill. Teaching a general strategy (Bruton & Samuda, 1981; Clarke & Nation, 1980; Williams, 1985) might also be effective because learners may become aware that integrating grammar information (part of speech of an unknown word) and discourse information (contextual clues) is important. Cloze exercises might be effective because they provide learners with the opportunity to do a lot of guessing, but might be effective only when learners know the majority of words in the context so that the target words may be guessable. Research might make it possible to identify where learners find difficulty and choose the appropriate teaching method that is most effective to improve their weaknesses.

The WPT will make it possible to investigate the effects of teaching on word part knowledge. A number of studies have been conducted with L1 children (Baumann, et al., 2002; Nagy, et al., 1993; Tyler & Nagy, 1989; White, Power, & White, 1989) indicating that teaching has an effect on knowledge of word parts. However, few attempts have been made to investigate the effects of teaching in the field of L2 acquisition. The WPT may contribute to determining the relative efficacy of teaching methods. Knowledge of word part form and meaning might effectively be gained by providing some example words with a particular affix rather than simply providing the relationships between affix form and meaning, because affixes do not exist on their own. Knowledge of word part use might be effectively gained when learners encounter sentences that include a word with a particular affix rather than encountering only an example word that includes the affix, because providing a sentence does not require learners to use metacognitive knowledge about part of speech.

Second, future research may also investigate the interrelationships among vocabulary size, word part knowledge, and guessing from context. This thesis has shown that knowledge of word parts is closely related to vocabulary size, but the relationship between the skill of guessing from context and vocabulary size remains to be investigated. Vocabulary size might be highly related to the guessing skill, because larger vocabulary size would allow learners to better comprehend the context and derive the meanings of unknown words more successfully which in turn provides learners with a greater chance of learning the meanings of unknown words. The combined effect of the guessing skill and word part knowledge on vocabulary size also remains to be investigated. It has been argued that guessing from context and word part knowledge play an important role in vocabulary learning, but little is known about the extent to

which each of them contributes to vocabulary size. This may be effectively investigated using the WPT and the GCT.

Third, the relationship between reading comprehension and word part knowledge may also be investigated. Positive evidence indicating a close relationship between them is well documented in L1 studies (Berninger, et al., 2010; Carlisle, 2000; Kuo & Anderson, 2006; Mahony, Singson, & Mann, 2000; Mahony, 1994; Nagy, et al., 2006; Singson, Mahony, & Mann, 2000; Tyler & Nagy, 1989). These studies indicate that word part knowledge contributes to the decoding of morphologically complex words and thus contributes to the development of reading comprehension. Research also indicates that word part knowledge is an increasingly important predictor of reading comprehension as children grow older. However, very few attempts have been made to investigate the relationship between reading comprehension and word part knowledge in the field of L2 acquisition. Qian (1999) found a significant correlation between the two (r=.64), but no significant contribution to reading comprehension was found in the regression model with the independent variables being vocabulary size, word association knowledge, and word part knowledge. As the WPT measures three aspects of word part knowledge, it may allow a more comprehensive approach towards examining the relationship between reading comprehension and word part knowledge.

Finally, creating standardised tests measuring other aspects of VLP such as the skill of dictionary use and phonological knowledge may be useful both for research and education. Future research might indicate that some aspects of VLP are more important to a group of learners with a particular level than others. Having a variety of VLP tests may be useful to teachers because they can diagnose their learners' weaknesses and indicate how they can become a more successful vocabulary learner.

7.4 Implications for Learning and Teaching

The GCT and the WPT may be of great practical use to teachers and learners because they clearly indicate learners' weaknesses in the skill of guessing from context and word part knowledge. Chapters 4 and 6 individually discussed ways in which the scores from the GCT and the WPT might be interpreted and reported to learners.

These two tests may also raise learners' awareness of vocabulary learning strategies. By taking the GCT, learners may become aware that guessing may be facilitated with knowledge of part of speech and contextual clues. By taking the WPT, learners may become aware that knowing a word part involves knowing its form, meaning, and use. Word part forms are worth explicit attention because L2 learners often have trouble with word forms especially when words share similar forms (Bensoussan & Laufer, 1984; Laufer, 1988). Some word parts have similar written forms such as *be-/de-* and *-ess/-ness*. Knowledge of word part meaning needs to be gained given that the first step in vocabulary learning is to establish form-meaning relationships (e.g., Nation, 2001; Schmitt, 2008). Knowledge of word part use is also important because some word parts (e.g., *-ness* in *kindness*) have the function of changing the part of speech of word stems but have hardly any substantial meaning in which case establishing a form-use link may be more practical than a form-meaning link.

In order to derive more accurate meanings of unknown words in authentic language use, it should be important to integrate the two sources of information (information from context and word parts) instead of relying too heavily on either of them. The use of the GCT and the WPT may raise learners' awareness of the importance of both types of information when dealing with unknown words. This is practically important because previous studies (Mori, 2002; Paribakht & Wesche, 1999; Parry,

1997) indicate that learners use different strategies when approaching unknown words: some learners may prefer to deal with unknown words analytically without relying on the information in the context, while others may try to guess the meanings of unknown words based on the information within the context rather than the information within the word elements. Contextual clues alone are not reliable for deriving the accurate meanings of unknown words because context does not always provide enough information to guess meanings (Beck, et al., 1983; Schatz & Baldwin, 1986). An analysis of word parts alone is not entirely reliable either because it is sometimes misleading; for example, *mother* is not made from *moth* and *-er*. Taken together, the GCT and the WPT are useful in raising learners' awareness of effective strategies for vocabulary learning as well as diagnosing their weaknesses in vocabulary learning.

The GCT and the WPT may also serve as achievement tests. As both tests have two equivalent forms, one form may be used for identifying learners' weaknesses prior to the instruction. Based on the test results, teachers may help learners to improve their learners' weaknesses through particular teaching methods or learning activities. After the instruction, the other form may be administered to their learners in order to see whether they have improved the skill of guessing or knowledge of word parts. The information from the second form is useful to teachers for making decisions regarding whether or not they can move on to the next unit of instruction. It is also useful to teachers for making decisions regarding appropriate modifications in their teaching methods or learning activities.

Currently, the GCT and the WPT are available only in a paper-based format. This will allow teachers to administer the tests without any special equipment in the classroom such as computers and the Internet. The tests might be more useful if they

were written in a web-based format so that learners can take the tests at any time and can receive prompt feedback on their scores. Such online tests might allow teachers to add their own question items such as student ID and a class name so that teachers can identify their learners' scores. Another function might be that when the tests are completed the scores are automatically calculated and are reported to learners using a bar graph which would clearly indicate their weaknesses. Online tests might also have the function of recording the response time for each item so that teachers could identify learners who took the tests without thinking carefully (too short response time) or those who relied on external resources such as a dictionary (too long response time).

7.5 Concluding Remarks

The purpose of this thesis has been to create and validate tests of VLP. Previous studies have created and validated a number of vocabulary tests which typically focus on how many words are known or how well a word is known (Beglar, 2010; Beglar & Hunt, 1999; Nation, 1983, 1990, 2006; Nation & Beglar, 2007; Read, 1993, 1998; Schmitt, et al., 2001). These tests are of theoretical value in investigating how different aspects of vocabulary knowledge are interrelated and how vocabulary knowledge is related to other language skills such as reading and listening. They also provide learners with useful information on their current level of vocabulary knowledge and clearly indicate how many words are needed for achieving a particular goal. However, previous vocabulary tests do not indicate how learners can become a good vocabulary learner. This thesis has been one of the first attempts to create such tests. Since teachers have no time to teach low-frequency words in class, it is important to help learners become proficient in vocabulary learning strategies so that they can effectively continue with

vocabulary learning on their own. The GCT and the WPT are expected to be useful tools for improving learners' VLP.

REFERENCES

Abbott, M. (2000). Identifying reliable generalizations for spelling words: The importance of multilevel analysis. *The Elementary School Journal, 101*(2), 233-245.

Aborn, M., Rubenstein, H., & Sterling, T. D. (1959). Sources of contextual constraint upon words in sentences. *Journal of Experimental Psychology, 57*(3), 171-180.

Ahmed, M. O. (1989). Vocabulary learning strategies. In P. Meara (Ed.), *Beyond Words* (pp. 3-14). London: BAAL\CILT.

Aitchison, J. (1994). *Words in the Mind* (2nd ed.). Oxford: Blackwell.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: APA.

Ames, W. S. (1966). The development of a classification scheme of contextual aids. *Reading Research Quarterly, 2*(1), 57-82.

Ames, W. S. (1970). The use of classification schemes in teaching the use of contextual aids. *Journal of Reading, 14*(1), 5-8, 50.

Anderson, R. C., & Nagy, W. E. (1991). Word meanings. In R. Barr, M. L. Kamil, P. Mosenthal & P. D. Pearson (Eds.), *Handbook of Reading Research* (Vol. 2, pp. 690-724). New York: Longman.

Andrich, D. (1988). *Rasch Models for Measurement*. Beverly Hills, CA: Sage Publications.

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development Serial No. 238, 58*(10).

Arden-Close, C. (1993). NNS readers' strategies for inferring the meanings of unknown words. *Reading in a Foreign Language, 9*(2), 867-893.

Artley, A. S. (1943). Teaching word-meaning through context. *Elementary English Review, 20*(1), 68-74.

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

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

Baddeley, A. (1990). *Human Memory*. London: Lawrence Erlbaum Associates.

Bauer, L. (1983). *English Word Formation*. Cambridge: Cambridge University Press.

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

Baumann, J. F., Edwards, E. C., Font, G., Tereshinski, C. A., Kame'enui, E. J., & Olejnik, S. (2002). Teaching morphemic and contextual analysis to fifth-grade students. *Reading Research Quarterly, 37*(2), 150-176.

Beard, R. (1982). The plural as a lexical derivation. *Glossa, 16*, 133-148.

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *Elementary School Journal, 83*(3), 177-181.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and the university word level vocabulary tests. *Language Testing, 16*(2), 131-162.

Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading, 7*(1), 15-32.

Berninger, V. W., Abbott, R. D., Nagy, W., & Carlisle, J. (2010). Growth in phonological, orthographic, and morphological awareness in Grades 1 to 6. *Journal of Psycholinguistic Research, 39*(2), 141-163.

Bock, C. (1948). Prefixes and suffixes. *Classical Journal, 44*, 132-133.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (second ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Bowey, J. A. (1996). On the association between phonological memory and receptive vocabulary in five-year-olds. *Journal of Experimental Child Psychology, 63*(1), 44-78.

Bowey, J. A. (2001). Nonword repetition and young children's receptive vocabulary: A longitudinal study. *Applied Psycholinguistics, 22*(3), 441-469.

Bradley, D. C. (1979). Lexical representation of derivational relation. In M. Aronoff & M. L. Kean (Eds.), *Juncture*. Sarasota, CA: Anma Libri.

Bradley, L., & Huxford, L. (1994). Organising sound and letter patterns for spelling. In G. D. A. Brown & N. C. Ellis (Eds.), *The Handbook of Spelling: Theory, Process and Intervention* (pp. 425-439). Chichester, UK: John Wiley & Sons.

Brown, C. (1993). Factors affecting the acquisition of vocabulary. In T. Huckin, H. M. & C. J. (Eds.), *Second Language Reading and Vocabulary* (pp. 263-286). Norwood, N.J: Ablex.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language, 20*(2), 136-163.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296-322.

Bruck, M., Treiman, R., Caravolas, M., Genesee, F., & Cassar, M. (1998). Spelling skills of children in whole language and phonics classrooms. *Applied Psycholinguistics, 19*(4), 669-684.

Bruton, A., & Samuda, V. (1981). Guessing words. *Modern English Teacher, 8*(3), 18-21.

Byrne, B., & Fielding-Barnsley, R. (1995). Evaluation of a program to teach phonemic awareness to young children: A 2- and 3-year follow-up and a new preschool trial. *Journal of Educational Psychology, 87*(3), 488-503.

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing: An Interdisciplinary Journal, 12*(3), 169-190.

Carnine, D., Kameenui, E. J., & Coyle, G. (1984). Utilization of contextual information in determining the meaning of unfamiliar words. *Reading Research Quarterly, 19*(2), 188-204.

Carpay, J. A. M. (1974). Foreign-language teaching and meaningful learning: A Soviet Russian point of view. *ITL, 25-26*, 161-187.

Carroll, J. B. (1940). Knowledge of English roots and affixes as related to vocabulary and Latin study. *Journal of Educational Research, 34*(2), 102-111.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage Word Frequency Book*. New York: Houghton Mifflin, Boston American Heritage.
Carroll, J. B., & Sapon, S. (1959). *The Modern Language Aptitude Test*. San Antonio, TX: Psychological Corporation.
Carstairs-McCarthy, A. (2002). *An Introduction to English Morphology: Words and their Structure*. Edinburgh: Edinburgh University Press.
Carton, A. S. (1971). Inferencing: A process in using and learning language. In P. Pimsleur & T. Quinn (Eds.), *The Psychology of Second Language Learning* (pp. 45-58). Cambridge: Cambridge University Press.
Casalis, S., & Louis-Alexandre, M. (2000). Morphological analysis, phonological analysis and learning to read French. *Reading and Writing: An Interdisciplinary Journal, 12*(3), 303-335.
Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272.
Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary learning. *Developmental Psychology, 32*(5), 867-873.
Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal, 80*(2), 183-198.
Clarke, D. F., & Nation, I. S. P. (1980). Guessing the meanings of words from context: Strategy and techniques. *System, 8*(3), 211-220.
Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Science* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.
Cole, P., Beauvillain, C., & Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words: a differential frequency effect. *Journal of Memory and Language, 28,* 1-13.
Cooper, T. C. (1999). Processing of idioms by L2 learners of English. *TESOL Quarterly, 33*(2), 233-262.
Costin, F. (1970). The optimal number of alternatives in multiple choice achievement tests: some empirical evidence for a mathematical proof. *Educational and Psychological Measurement, 30*(2), 353-358.
Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement, 32*(4), 1035-1038.
Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.
Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology, 104*(3), 268-294.
Crookes, G., & Schmidt, R. W. (1991). Motivation: Reopening the research agenda. *Language Learning, 41*(4), 469-512.
Cutler, A., Hawkins, J. A., & Gilligan, G. (1985). The suffixing preference: A processing explanation. *Linguistics, 23*(5), 723-758.
Daulton, F. E. (2004). *Gairaigo -- the built-in lexicon?: The common loan words in Japanese based on high-frequency English vocabulary and their effect on language acquisition.* Unpublished doctoral dissertation, Victoria University of Wellington, Wellington.

Day, R. R., Omura, C., & Hiramatsu, M. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language, 7*(2), 541-551.

de Bot, K., Paribakht, T., & Wesche, M. (1997). Towards a lexical processing model for the study of second language vocabulary acquisition: Evidence from ESL reading. *Studies in Second Language Acquisition, 19*, 309-329.

de Jong, P. F., Seveke, M. J., & van Veen, M. (2000). Phonological sensitivity and the acquisition of new words in children. *Journal of Experimental Child Psychology, 76*(4), 275-301.

Deighton, L. C. (1959). *Vocabulary Development in the Classroom*. New York: Columbia University Press.

Draba, R. E. (1977). The identification and interpretation of item bias. *MESA Research Memorandum, No. 25*.

Dubin, F., & Olshtain, E. (1993). Predicting word meanings from contextual clues: Evidence from L1 readers. In T. Huckin, M. Haynes & J. Coady (Eds.), *Second Language Reading and Vocabulary* (pp. 181-202). Norwood, N.J.: Ablex.

Dulin, K. L. (1969). New research on context clues. *Journal of Reading, 13*(1), 33-38, 53.

Dulin, K. L. (1970). Using context clues in word recognition and comprehension. *Reading Teacher, 23*(5), 440-445.

Dupuy, B., & Krashen, S. D. (1993). Incidental vocabulary acquisition in French as a foreign language. *Applied Language Learning, 4*(1&2), 55-63.

Educational Testing Service. (1987). *Reading for TOEFL: Workbook*. Princeton, NJ: Educational Testing Service.

Ehrman, M. E., & Oxford, R. L. (1995). Cognition plus: Correlates of language learning success. *Modern Language Journal, 79*(1), 67-89.

Elgort, I. (2007). *The role of intentional decontextualised learning in second language vocabulary acquisition: Evidence from primed lexical decision tasks with advanced bilinguals* Unpublished doctoral dissertation, Victoria University of Wellington, New Zealand.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.

Ellis, N. C., & Sinclair, S. G. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *Quarterly Journal of Experimental Psychology, 49A*(1), 234-250.

Ellis, R. (1990). *Instructed Second Language Acquisition*. Oxford: Basil Blackwell.

Embretson, S. E., & Hershberger, S. L. (1999). The new rules of measurement: what every psychologist and educator should know. In S. E. Embretson & S. L. Hershberger (Eds.), *Summary and Future of Psychometric Methods in Testing* (pp. 243-254). Mahwah, NJ: Lawrence Erlbaum Associates.

Fisher Jr., W. P. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective Measurement: Theory into Practice* (Vol. 2). Norwood, NJ: Ablex.

Fowler, A. E. (1991). How early phonological development might set the stage for phoneme awareness. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological Processes in Literacy* (pp. 97-117). Hillsdale, NJ: Erlbaum.

Fraser, C. A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition, 21*, 225-241.

Freyd, P., & Baron, J. (1982). Individual differences in acquisition of derivational morphology. *Journal of Verbal Learning and Verbal Behavior, 21*(3), 282-295.

Fukkink, R., Blok, H., & de Glopper, K. (2001). Deriving word meaning from written context: A multicomponential skill. *Language Learning, 51*(3), 477-496.

Fukkink, R. G., & de Glopper, K. (1998). Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research, 68*(4), 450-469.

Gardner, R. C., & MacIntyre, P. D. (1991). An instrumental motivation in language study: Who said it isn't effective? *Studies in Second Language Acquisition, 13*(1), 57-72.

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory and Cognition, 23*(1), 83-94.

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28*, 200-213.

Gathercole, S. E., Service, E., Hitch, G. J., Adams, A. M., & Martin, A. J. (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Applied Cognitive Psychology, 13*(1), 65-77.

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

Graves, M. (1984). Selecting vocabulary to teach in the intermediate and secondary grades. In J. Flood (Ed.), *Promoting Reading Comprehension* (pp. 245-260). Newark, DE: International Reading Association.

Gu, Y., & Johnson, R. K. (1996). Vocabulary learning strategies and language learning outcomes. *Language Learning, 46*(4), 643-679.

Haastrup, K. (1985). Lexical inferencing - a study of procedures in reception. *Scandinavian Working Papers on Bilingualism, 5*, 63-87.

Haastrup, K. (1987). Using thinking aloud and retrospection to uncover learners' lexical inferencing procedures. In C. Faerch & G. Kasper (Eds.), *Introspection in Second Language Research* (pp. 197-212). Clevedon: Multilingual Matters.

Haastrup, K. (1991). *Lexical Inferencing Procedures or Talking about Words*. Tubingen: Gunter Narr.

Harwood, F. W., & Wright, A. M. (1956). Statistical study of English word formation. *Language, 32*, 260-273.

Haynes, M. (1993). Patterns and perils of guessing in second language reading. In T. Huckin, M. Haynes & J. Coady (Eds.), *Second Language Reading and Vocabulary* (pp. 46-64). Norwood, N.J.: Ablex.

Hill, M., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *IRAL, 41*, 87-106.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689-696.

Honeyfield, J. (1977). Simplification. *TESOL Quarterly, 11*(4), 431-440.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207-223.

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

Huckin, T., & Bloch, J. (1993). Strategies for inferring word meanings: A cognitive model. In T. Huckin, M. Haynes & J. Coady (Eds.), *Second Language Reading and Vocabulary* (pp. 153-178). Norwood, N.J.: Ablex.

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language, 30*, 685-701.

Hulstijn, J., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *Modern Language Journal, 80*(3), 327-339.

Hulstijn, J., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539-558.

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 113-125). London: Macmillan.

Ishii, T., & Schmitt, N. (2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal, 40*(1), 5-22.

Jenkins, J. R., & Dixon, R. (1983). Vocabulary learning. *Contemporary Educational Psychology, 8*, 237-260.

Jenkins, J. R., Matlock, B., & Slocum, T. A. (1989). Two approaches to vocabulary instruction: the teaching of individual word meanings and practice in deriving word meanings from context. *Reading Research Quarterly, 24*(2), 215-235.

Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal, 21*(4), 767-787.

Jiang, N. (2004). Semantic transfer and development in adult L2 vocabulary acquisition. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language: Selection, Acquisition, and Testing* (pp. 101-126). Amsterdam: John Benjamins.

Joe, A. (1995). Text-based tasks and incidental vocabulary learning. *Second Language Research, 11*(2), 149-158.

Johnson, D., & Pearson, P. D. (1984). *Teaching Reading Vocabulary*. New York: Holt, Rinehart & Winston.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*, 152-176.

Kim, Y. (2011). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 61*(1), 100-140.

Knight, S. M. (1994). Dictionary use while reading: the effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal, 78*(3), 285-299.

Kolstad, R. K., Briggs, L. D., & Kolstad, R. A. (1985). Multiple-choice classroom achievement tests: Performance on items with five vs. three choices. *College Student Journal, 19*(4), 427-431.

Kolstad, R. K., Kolstad, R. A., & Wagner, M. J. (1986). Performance on 3-choice versus 5-choice MC items that measure different skills. *Educational Research Quarterly, 10*, 4-8.

Kuhn, M. R., & Stahl, S. A. (1998). Teaching children to learn word meanings from context. *Journal of Literacy Research, 30*(1), 119-138.

Kuo, L., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist, 41*(3), 161-180.

Laufer, B. (1988). The concept of `synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education, 2*(2), 113-132.
Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. L. a. M. Nordman (Ed.), *Special Language: From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters.
Laufer, B. (1991). The development of L2 lexis in the expression of the advanced language learner. *Modern Language Journal, 75*(4), 440-448.
Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 126-132). London: Macmillan.
Laufer, B. (1997). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 140-155). Cambridge: Cambridge University Press.
Laufer, B., & Bensoussan, M. (1982). Meaning is in the eye of the beholder. *English Teaching Forum, 20*(2), 10-13.
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.
Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.
Laufer, B., & Hill, M. (2000). What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? *Language Learning and Technology, 3*(2), 58-76.
Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1-26.
Laufer, B., & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing, 16*(1), 36-55.
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.
Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal, 28*(1), 89-108.
Laufer, B., & Sim, D. D. (1985). Taking the easy way out: Non-use and misuse of clues in EFL reading. *English Teaching Forum, 23*(2), 7-10, 20.
Lawson, M. J., & Hogben, D. (1996). The vocabulary-learning strategies of foreign-language students. *Language Learning, 46*(1), 101-135.
Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. Harlow: Longman.
Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.
Linacre, J. M. (1995). Prioritizing misfit indicators. *Rasch Measurement Transactions, 9*(2), 422-423.
Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.
Linacre, J. M. (2003). Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions, 17*(1), 918.

Linacre, J. M. (2010a). *A User's Guide to WINSTEPS® MINISTEPS Rasch-Model Computer Programs: Program Manual 3.70.0*. Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2010b). WINSTEPS (Version 3.70.0) Rasch measurement computer program. Beaverton, Oregon: Winsteps.com.

Linacre, J. M., & Tennant, A. (2009). More about critical eigenvalue sizes in standardized-residual principal components analysis (PCA). *Rasch Measurement Transactions, 23*(3), 1228.

Liu, N., & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal, 16*(1), 33-42.

Lundberg, I., Frost, J., & Peterson, O. P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23*(3), 263-284.

Luppescu, S., & Day, R. R. (1993). Reading, dictionaries and vocabulary learning. *Language Learning, 43*(2), 263-287.

Mahony, D., Singson, M., & Mann, V. (2000). Reading ability and sensitivity to morphological relations. *Reading and Writing: An Interdisciplinary Journal, 12*(3), 191-218.

Mahony, D. L. (1994). Using sensitivity to word structure to explain variance in high school and college level reading ability. *Reading and Writing: An Interdisciplinary Journal, 6*(1), 19-44.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Marchand, H. (1960). *The Categories and Types of Present-Day English Word-Formation*. Gesamtherstellung: Hubert & Co.

Masoura, E. V., & Gathercole, S. E. (1999). Phonological short-term memory and foreign language learning. *International Journal of Psychology, 34*(5/6), 383-388.

Masoura, E. V., & Gathercole, S. E. (2005). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory, 13*(3/4), 422-429.

McCarthy, M. (1990). *Vocabulary*. Oxford: Oxford University Press.

McCullough, C. M. (1943). Learning to use context clues. *Elementary English Review, 20*, 140-143.

McCullough, C. M. (1945). The recognition of context clues in reading. *Elementary English Review, 22*(1), 1-5.

McCullough, C. M. (1958). Context aids in reading. *Reading Teacher, 11*(4), 225-229.

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

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*(2), 142-151.

Meara, P., & Jones, G. (1990). *Eurocentres Vocabulary Size Test. 10KA*. Zurich: Eurocentres.

Medical Outcomes Trust Scientific Advisory Committee. (1995). Instrument Review Criteria. *Medical Outcomes Trust Bulletin*, 1-4.

Meng, X.-L., Rosenthal, R., & Rubin, D., B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*(1), 172-175.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Procedures to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word Recognition in Beginning Literacy* (pp. 89-120). Mahwah, NJ: Erlbaum.

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

Mochizuki, M. (1998). Nihonjin eigo gakushusha no setsuji rikai chosa [Understanding English affixes by Japanese learners. *Reitaku Review, 4*, 100-120.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System, 28*(2), 291-304.

Moir, J., & Nation, I. S. P. (2002). Learners' use of strategies for effective vocabulary learning. *Prospect, 17*(1), 15-35.

Mori, Y. (2002). Individual differences in the itntegration of information from context and word parts in interpreting unknown kanji words. *Applied Psycholinguistics, 23*(3), 375-397.

Mori, Y., & Nagy, W. (1999). Integration of information from context and word elements in interpreting novel kanji compounds. *Reading Research Quarterly, 34*(1), 80-101.

Morrison, L. (1996). Talking about words: A study of French as a second language learners' lexical inferencing procedures. *Canadian Modern Language Review, 53*(1), 41-75.

Nagy, W., Berninger, V., Abbott, R., Vaughan, K., & Vermeulen, K. (2003). Relationship of morphology and other language skills to literacy skills in at-risk second graders and at-risk fourth grade writers. *Journal of Educational Psychology, 95*(4), 739-742.

Nagy, W., Berninger, V. W., & Abbott, R. D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of Educational Psychology, 98*(1), 134-147.

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.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly, 19*(3), 304-330.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal, 24*(2), 237-270.

Nagy, W. E., Diakidoy, I. N., & Anderson, R. C. (1993). The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives. *Journal of Reading Behavior, 25*(2), 155-169.

Nagy, W. E., Herman, P., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*(2), 233-253.

Nassaji, H. (2003). L2 vocabulary learning from context: strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly, 37*(4), 645-670.

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

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. Rowley, Mass.: Newbury House.

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. (2009). *Teaching ESL/EFL Reading and Writing*. New York: Routledge.

Nation, I. S. P., & Coady, J. (1988). Vocabulary and reading. In R. Carter & M. McCarthy (Eds.), *Vocabulary and Language Teaching* (pp. 97-110). London: Longman.

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

Oxford, R., & Crookall, D. (1990). Vocabulary learning: A critical analysis of techniques. *TESL Canada Journal, 7*(2), 9-30.

Papagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language, 30*, 331-347.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary development. In J. Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 174-200). New York: Cambridge University Press.

Paribakht, T. S., & Wesche, M. (1999). Reading and "incidental" L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language Acquisition, 21*, 195-224.

Parry, K. (1991). Building a vocabulary through academic reading. *TESOL Quarterly, 25*(4), 629-653.

Parry, K. (1997). Vocabulary and comprehension: Two portraits. In J. Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition* (pp. 55-68). Cambridge: Cambridge University Press.

Peters, E. (2007). Manipulating L2 learners' online dictionary use and its effect on L2 word retention. *Language Learning and Technology, 11*(2), 36-58.

Pitts, M., White, H., & Krashen, S. (1989). Acquiring second language vocabulary through reading: A replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language, 5*(2), 271-275.

Prince, P. (1996). Second language vocabulary learning: The role of context versus translations as a function of proficiency. *Modern Language Journal, 80*(4), 478-493.
Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprenhension. *Canadian Modern Language Review, 56*(2), 282-307.
Quealy, R. J. (1969). Senior high school students' use of context aids in reading. *Reading Research Quarterly, 4*(4), 512-532.
Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions, 19*(1), 1012.
Rankin, E. F., & Overholser, B. M. (1969). Reaction of intermediate grade children to contextual clues. *Journal of Reading Behavior, 1*(3), 50-73.
Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.
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, N.J.: Lawrence Erlbaum Associates.
Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.
Read, J., & Chapelle, C. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 3-32.
Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*(1), 77-89.
Roberts, T. A., & Meiring, A. (2006). Teaching phonics in the context of children's literature or spelling: Influences on first-grade reading, spelling and writing and fifth-grade comprehension. *Journal of Educational Psychology, 98*(4), 690-713.
Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13.
Rosenthal, J., & Ehri, L. C. (2008). The mnemonic value of orthography for vocabulary learning. *Journal of Educational Psychology, 100*(1), 173-191.
Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabuary acquisition through reading. *Studies in Second Language Acquisition, 21*(1), 589-619.
Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System, 6*(2), 72-78.
Schatz, E. K., & Baldwin, R. S. (1986). Context clues are unreliable predictors of word meaning. *Reading Research Quarterly, 21*(4), 439-453.
Schmidt, R. W., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portugese. In R. Day (Ed.), *Talking to Learn: Conversation in Second Language Acquisition*. Rowley, Mass.: Newbury House.
Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 199-227). Cambridge: Cambridge University Press.
Schmitt, N. (1998). Tracking the incidental 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.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 325-363.

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.

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.

Schmitt, N., & Zimmerman, C. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36*(2), 145-171.

Scholfield, P. J. (1982). Using the English dictionary for comprehension. *TESOL Quarterly, 16*(2), 185-194.

Schumacker, R. E., & Muchinsky, P. M. (1996). Disattenuating correlation coefficients. *Rasch Measurement Transactions, 10*(1), 479.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., & Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology, 62*(3), 288-295.

Segui, J., & Zubizaretta, M. L. (1985). Mental representation of morphologically complex words and lexical access. *Linguistics, 23*(5), 759-774.

Seibert, L. C. (1945). A study on the practice of guessing word meanings from a context. *Modern Language Journal, 29*(4), 296-323.

Service, E. (1992). Phonology, working memory, and foreign language learning. *Quarterly Journal of Experimental Psychology, 45A (1)*, 21-50.

Shu, H., Anderson, R. C., & Zhang, Z. (1995). Incidental learning of word meanings while reading: A Chinese and American cross-cultural study. *Reading Research Quarterly, 30*(1), 76-95.

Singson, M., Mahony, D., & Mann, V. (2000). The relation between reading ability and morphological skills: Evidence from derivational suffixes. *Reading and Writing: An Interdisciplinary Journal, 12*(3), 219-252.

Slinde, J. A., & Linn, R. L. (1979). The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement, 3*(4), 437-452.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*(33), 1-11.

Smith Jr., E. V. (2004a). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch Measurement: Theory, Models and Applications* (pp. 575-600). Maple Grove, Minnesota: JAM Press.

Smith Jr., E. V. (2004b). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch Measurement: Theory, Models and Applications* (pp. 93-122). Maple Grove, Minnesota: JAM Press.

Smith Jr., E. V. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement, 6*, 147-163.

Smith, R. M. (1992). *Applications of Rasch Measurement*. Chicago: MESA Press.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3*(1), 25-40.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1*, 199-218.

Smith, R. M., & Miao, C. (1994). Assessing dimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective Measurement: Theory into Practice* (Vol. 2, pp. 316-327). Norwood, NJ: Ablex.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

Smith, R. M., & Suh, K. K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates. *Journal of Applied Measurement, 4*(2), 153-163.

Snowling, M., Goulandris, N., Bowlby, M., & Howell, P. (1986). Segmentation and speech perception in relation to reading skill: a developmental analysis. *Journal of Experimental Child Psychology, 41*(3), 489-507.

Spache, G., & Berg, P. (1955). *The Art of Efficient Reading*. New York: Macmillan.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72-101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271-295.

Stauffer, R. G. (1942). A study of prefixes in the Thorndike list to establish a list of prefixes that should be taught in the elementary school. *Journal of Educational Research, 35*(6), 453-458.

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. McKeown & M. Curtis (Eds.), *The Nature of Vocabulary Acquisition* (pp. 89-105). Mahwah, N.J. : Lawrence Erlbaum Associates.

Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. *American Psychologist, 38*, 878-893.

Stevens, J. (2002). *Applied Multivariate Statistics for the Social Sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.

Strang, R. M. (1944). How students attack unfamiliar words. *The English Journal, 33*(2), 88-93.

Taft, M. (1979). Recognition of affixed wordsand the word frequency effect. *Memory & Cognition, 7*(4), 263-272.

Taft, M. (1985). The decoding of words in lexical access: A review of the morphographic approach. In D. Benser, T. G. Waller & Mackinnon (Eds.), *Reading Research: Advances in Theory and Practice*. London: Academic Press.

Taft, M., & Forster, K. I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior, 15*(6), 607-620.

Thorndike, E. L. (1932). *Teacher's Word Book of 20,000 Words*. New York: Teachers College Columbia.

Thorndike, E. L. (1941). *The Teaching of English Suffixes*. New York: Teachers College, Columbia University.
Thorndike, E. L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College Columbia University.
Tono, Y. (1988). Assessment of EFL learners' dictionary using skills. *JACET Bulletin, 19*, 103-126.
Tristan, A. (2006). An adjustment for sample size in DIF analysis. *Rasch Measurement Transactions, 20*(2), 1070-1071.
Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language, 28*, 649-667.
Tyler, A., & Nagy, W. (1990). Use of derivational morphology during reading. *Cognition, 36*, 17-34.
van Parreren, C. F. (1975). First and second-language learning compared. In A. J. van Essen & J. P. Menting (Eds.), *The Context of Foreign-Language Learning* (pp. 100-116). Assen: Van Gorcum.
Walley, A. C. (1993). The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review, 13*(3), 286-350.
Walters, J. (2004). Teaching the use of context to infer meaning: A longitudinal survey of L1 and L2 vocabulary research. *Language Teaching, 37*, 243-252.
Walters, J. (2006). Methods of teaching inferring meaning from context. *RELC Journal, 37*(2), 176-190.
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*(2), 130-163.
Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*(1), 33-52.
Webb, S. (2007a). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46-65.
Webb, S. (2007b). The effects of synonymy on second-language vocabulary learning. *Reading in a Foreign Language, 19*(2), 120-136.
Webb, S. (2007c). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research, 11*(1), 63-81.
Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language, 20*(2), 232-245.
Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal, 40*(3), 360-376.
Weitzman, R. A. (1996). The Rasch model plus guessing. *Educational and Psychological Measurement, 56*(5), 779-790.
Wesche, M., & Paribakht , T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*(1), 13-40.
White, T. G., Power, M. A., & White, S. (1989). Morphological analysis: Implications for teaching and understanding vocabulary growth. *Reading Research Quarterly, 24*(3), 283-304.
Williams, R. (1985). Teaching vocabulary recognition strategies in ESP reading. *ESP Journal, 4*(2), 121-131.

Wolfe, E. W., & Smith Jr., E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part 2 - Validation activities. *Journal of Applied Measurement, 8*, 204-234.
Wright, B. D. (1992). IRT in the 1990s: Which models work best? *Rasch Measurement Transactions, 6*(1), 196-200.
Wright, B. D. (1995). 3PL or Rasch? *Rasch Measurement Transactions, 9*(1), 408-409.
Wright, B. D., & Douglas, G. A. (1976). Rasch item analysis by hand. *MESA Research Memorandum, No. 21*.
Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.
Wysocki, K., & Jenkins, J. R. (1987). Deriving word meanings through morphological generalization. *Reading Research Quarterly, 22*(1), 66-81.
Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language comprehension. *Language Learning, 44*(2), 189-219.

Appendix A. Test words, nonsense words, part of speech, context clues and place

Item	Test word	Nonsense word	PoS	Discouse clue	Place
1	ostensibly	secomantly	Adv	Contrast/comparison	Inside
2	damsel	tanave	Noun	Contrast/comparison	Inside
3	abseil	drunge	Verb	Contrast/comparison	Inside
4	dispensable	fedensable	Adj	Contrast/comparison	Outside
5	homicide	hurblige	Noun	Contrast/comparison	Outside
6	preen	climp	Verb	Synonym	Inside
7	blemish	widonce	Noun	Synonym	Inside
8	annotate	ceredate	Verb	Synonym	Inside
9	thicket	burtint	Noun	Synonym	Outside
10	abscond(ed)	turmilted	Verb	Synonym	Outside
11	conjecture	melabosure	Noun	Direct description	Inside
12	hutch	rotep	Noun	Direct description	Inside
13	abound	vanink	Verb	Direct description	Inside
14	scant	debin	Adj	Direct description	Outside
15	disparate	tengerate	Adj	Direct description	Outside
16	enigma	botile	Noun	Appositive	Inside
17	mutable	nogable	Adj	Appositive	Inside
18	impassively	monsitively	Adv	Appositive	Inside
19	indolent	serident	Adj	Appositive	Inside
20	amnesia	tarrand	Noun	Appositive	Inside
21	gabble(d)	blurged	Verb	Indirect description	Inside
22	zenith	liatom	Noun	Indirect description	Inside
23	platter	crannel	Noun	Indirect description	Inside
24	tremor	vansel	Noun	Indirect description	Outside
25	rumple(d)	ceacled	Verb	Indirect description	Outside
26	twirl	gorel	Noun	Cause/effect	Inside
27	languish(ed)	blonounded	Verb	Cause/effect	Inside
28	preponderance	mordontance	Noun	Cause/effect	Inside
29	morosely	ronditely	Adv	Cause/effect	Outside
30	ophthalmic	strocastic	Adj	Cause/effect	Outside
31	loll(ed)	bloyed	Verb	Restatement	Outside
32	rundown	sharrel	Noun	Restatement	Outside
33	clandestine	devertine	Adj	Restatement	Outside
34	feign(ed)	smanted	Verb	Restatement	Outside
35	cacophony	strantony	Noun	Restatement	Outside
36	connoisseur	candintock	Noun	Modification	Inside
37	refectory	bempurstory	Noun	Modification	Inside
38	torpor	tarint	Noun	Modification	Inside
39	propulsion	contression	Noun	Modification	Inside
40	aperture	gosposure	Noun	Modification	Inside
41	scion	nadge	Noun	Reference	Inside
42	appease	drumple	Verb	Reference	Inside
43	facile	fentile	Adj	Reference	Inside
44	trilby	rotice	Noun	Reference	Outside
45	stagnant	rubidant	Adj	Reference	Outside
46	extricate	densodate	Verb	Words in series	Inside

Item	Test word	Nonsense word	PoS	Discouse clue	Place
47	astern	ascrice	Adv	Words in series	Inside
48	ravenously	ferduously	Adv	Words in series	Inside
49	jocular	dacular	Adj	Words in series	Inside
50	disembark(ed)	diffuntled	Verb	Words in series	Inside
51	conundrum	scanegeon	Noun	Association	Inside
52	seabed	roocle	Noun	Association	Inside
53	avert(ed)	chonked	Verb	Association	Inside
54	encase(d)	wincled	Verb	Association	Inside
55	singe(d)	famped	Verb	Association	Inside
56	autopsy	sparbon	Noun	Example	Inside
57	beverage(s)	duterages	Noun	Example	Inside
58	anaesthesia	delincert	Noun	Example	Inside
59	orthographically	decontanically	Adv	Example	Outside
60	venerate(d)	mericated	Verb	Example	Outside

Appendix B. List of affixes

No.	Affix	M	U	No.	Affix	M	U	No.	Affix	M	U
1	a- (toward)	✔	✔	41	un-	✔		81	-i	✔	
2	a- (not)	✔		42	uni-	✔		82	-ian	✔	
3	ab-	✔		43	-able	✔	✔	83	-ible	✔	✔
4	anti-	✔		44	-age		✔	84	-ic		✔
5	arch-	✔		45	-al (a)		✔	85	-ify		✔
6	auto-	✔		46	-al (n)		✔	86	-ion		✔
7	be-		✔	47	-an	✔		87	-ise		✔
8	bi-	✔		48	-ance		✔	88	-ish		✔
9	circum-	✔		49	-ancy		✔	89	-ism	✔	
10	co-	✔		50	-ant (a)		✔	90	-ist	✔	
11	counter-	✔		51	-ant (n)	✔	✔	91	-ite	✔	
12	de-	✔		52	-ar		✔	92	-ition		✔
13	dis-	✔		53	-ary (a)		✔	93	-ity		✔
14	em-		✔	54	-ary (n)	✔	✔	94	-ive		✔
15	en-		✔	55	-ate (a)		✔	95	-ize		✔
16	ex-	✔		56	-ate (v)		✔	96	-less	✔	✔
17	fore-	✔		57	-atic		✔	97	-let	✔	
18	hyper-	✔		58	-ation		✔	98	-ling	✔	
19	il-	✔		59	-atory		✔	99	-ly (a)		✔
20	im-	✔		60	-cy		✔	100	-ly (adv)		✔
21	in-	✔		61	-dom	✔		101	-ment		✔
22	inter-	✔		62	-ee	✔	✔	102	-most	✔	
23	ir-	✔		63	-eer	✔		103	-ness		✔
24	mal-	✔		64	-en (a)	✔	✔	104	-or	✔	✔
25	micro-	✔		65	-en (v)		✔	105	-ory		✔
26	mid-	✔		66	-ence		✔	106	-ous		✔
27	mis-	✔		67	-ency		✔	107	-ship	✔	
28	mono-	✔		68	-ent (a)		✔	108	-some		✔
29	multi-	✔		69	-ent (n)	✔	✔	109	-ster	✔	
30	neo-	✔		70	-er	✔	✔	110	-th (n)		✔
31	non-	✔		71	-ery		✔	111	-th (ordinal)	✔	
32	post-	✔		72	-ese	✔		112	-ty		✔
33	pre-	✔		73	-esque		✔	113	-ure		✔
34	pro-	✔		74	-ess	✔		114	-ward	✔	✔
35	re-	✔		75	-et	✔		115	-ways	✔	✔
36	semi-	✔		76	-ette	✔		116	-wise	✔	✔
37	sub-	✔		77	-fold	✔		117	-y (a)		✔
38	super-	✔		78	-ful (a)		✔	118	-y (n)		✔
39	sur-	✔		79	-ful (n)	✔					
40	trans-	✔		80	-hood	✔					

Note: The form section measures all affixes. “M” = meaning section, “U” = use section.

Appendix C. Affixes not included in the WPT

Study	Affixes
Bock (1948)	Prefixes: *ad-* (adjoin), *ambi-* (ambiguous), *ante-* (antenatal), *bene-* (benefaction), *com-* (combine), *contra-* (contradict), *equi-* (equidistance), *ex-* (export), *extra-* (extraordinary), *in-* (include), *intra-* (intramural), *juxta-* (juxtapose), *ob-* (obstruct), *per-* (percolate), *quadra-* (quadrangle), *retro-* (retrospect), *satis-* (satisfy), *sine-* (?), *subter-* (subterfuge), *ultra-* (ultraviolet), *vari-* (variometer), *vice-* (vice-president). Suffixes: *-ade* (lemonade), *-ain* (?), *-arian* (librarian), *-arium* (aquarium), *-e* (?), *-esy* (?), *-fic/-fice* (specific), *-fix* (?), *-ice/-ix* (justice), *-icle* (article), *-ide* (chloride),*-ile* (percentile), *-ine* (heroine), *-late* (?), *-mony* (ceremony), *-orium* (auditorium), *-tine* (?), *-tude* (magnitude), *-ule* (module), *-uscle* (?).
Bauer & Nation (1993)	*ante-* (antenatal).
Carroll (1940)	Prefixes: *ad-*, *com-*, *con-*, *di-*, *dia-*, *ex-*, *e-*, *extra-*, *in-* (into), *im-*, *per-* . Suffixes: *-er* (comparative), *-tude*.
Freyd & Baron (1982)	*-ed* (disordered), *-itude* (servitude).
Mochizuki (1998)	Prefixes: *ambi-* (ambidextrous), *ana-* (anachronism), *com-* (combine), *contra-* (contradict), *extra-* (extracurricular), *over-* (overwork), *under-* (understatement). Suffixes: *-ed* (red-headed).
Mochizuki & Aizawa (2000)	*ex-* (export).
Nation (2001)	*ad-* (advert), *ante-* (antenatal), *com-* (combine), *ex-* (exclude), *in-* (include), *ob-* (obstruct), *per-* (percolate), *pro-* (proceed).
Schmitt & Meara (1997)	*-ed* (agreed), *-ing* (agreeing), *-s* (agrees).
Stauffer (1942)	*ad-* (admit), *ambi-* (ambiguous), *amphi-* (amphibian), *an-* (anarch), *ana-* (anatomy), *ante-* (antedote), *apo-* (apostasy), *bene-* (benefactor), *cata-* (catalog), *com-* (commemorate), *contra-* (contradict), *di-* (dilemma), *dia-* (diagnose), *dys-* (dysentery), *ec-* (eccentric), *enter-* (enterprise), *epi-* (epigram), *equi-* (equidistant), *ex-* (exaggerate), *extra-* (extraordinary), *for-* (forbid), *hemi-* (hemisphere), *hypo-* (hypothenuse), *in-* (incarnate), *meta-* (metamorphosis), *ob-* (obstacle), *off-* (offset), *para-* (paraphrase), *per-* (percolate), *peri-* (periphery), *poly-* (polysyllable), *pro-* (proceed), *retro-* (retrospect), *se-* (secede), *syn-* (syncopate), *tri-* (triangle).
Tyler and Nagy (1989)	*-like* (childlike), *-s* (books).
Wysocki & Jenkins (1987)	*-ing*.