An investigation into metaphoric competence in the L2:

A linguistic approach

David Malcolm O'Reilly

PhD

University of York

Education

October 2017

Abstract

Within the field of L2 metaphoric competence (MC) research, Low's (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skills and (sub)competences have existed for 29 and 11 years respectively, but have never been elicited or used to develop tests. Consequently, the extent to which they are underpinned by more fundamental (sub)constructs is unclear. With a few exceptions (e.g., Littlemore, 2001), L2 MC tests to date have been limited in scope (e.g., Aleshtar & Dowlatabadi, 2014; Azuma, 2005; Hashemian & Nezhad, 2007; Zhao, Yu, & Yang, 2014). Available research shows that L2 MC correlates with L2 vocabulary knowledge and proficiency (Aleshtar & Dowlatabadi, 2014; Zhao et al., 2014), but negligibly with time spent in an L2 immersion setting (Azuma, 2005). However, the ability of these measures to predict L2 MC is unknown, as is the change in the receptive/productive correlation strength as L2 proficiency increases.

In response to these gaps, a large battery of L2 MC tests aimed at eliciting Low's (1988) and Littlemore and Low's (2006a, 2006b) constructs was developed and administered to 112 NNSs of English (L1 Chinese) and 31 English NSs, along with vocabulary knowledge and (NNSs only) general proficiency tests. Data cleaning showed inherent, operationalisation problems. Exploratory Factor Analysis revealed four metaphor-related factors, with MANOVA and independent samples *t*-tests showing statistical NNS and NS differences for only one of these: English Grammatical Metaphoric Competence. Multiple regression revealed that the Oxford Online Placement Test best predicted L2 receptive MC, whereas L2 vocabulary depth measured by the Word Associates Test (Read, 1998) best predicted L2 productive MC. Time spent living in the UK had no predictive power, and the receptive/productive correlation weakened with increased L2 proficiency. Implications for theory, test development, the transferability of models and predictors (e.g., to NNSs with other L1s) and EFL teaching are discussed.

Table of Contents

List of tables

List of figures

Conventions

The following conventions have been used throughout this thesis.

Invented example/linguistic metaphor	'the White House issued a statement'
Conceptual metaphor	THE BODY IS A CONTAINER FOR THE EMOTIONS
Conceptual domain (used on its own)	THE MIND
Question that the reader is invited to ask	One might well ask, *what do the authors mean?*
Referring to a test item	'don't worry…go out and break a leg. In fact, go out and_____' (item 4)
Underlining to highlight the metaphorical part of an utterance	…you said you would think about the Prime Ministership <u>if the ball came loose from the scrum…</u>

Acknowledgements

I would like to express my sincerest thanks to my supervisor, Dr Emma Marsden. Your expertise, guidance and support have been invaluable. I am also deeply grateful to Dr Graham Low for introducing a fresh-faced MA student to the world of metaphor and believing he had the right mind for a PhD, and to Dr Bill Soden for his time and expertise. Thanks to all the Chinese and British participants who took part in this study, without you it would not have been possible. I also thank Gillian O'Reilly and Sophie Thompson for their diligence and assistance with scoring, and Dr Nora McIntyre for vital training and advice. I thank Gillian again and my father Alan and brother Michael for everything they have given me. Finally, I express my love and gratitude to Jelena O'Reilly (née Horvatić), who I met on the first day of the PhD, and to whom I am now married. You make everything better.

Author's declaration

I declare that this thesis, including all data presented in it, is original work and that I am the sole author. This work has not previously been presented for an award at this, or any other, University. All data collection and analysis was carried out by the author, David O'Reilly, with the assistance of Gillian O'Reilly and Sophie Thompson, who helped score responses to limited production questions in the MC Test Battery. All sources used are acknowledged as References.

Chapter 1: Introduction

1.1 The research context

Research into how people comprehend and produce metaphor (i.e., their metaphoric competence) has been around for several decades. The first such studies measured metaphoric competence in adult native (L1) speakers rather than second or foreign (L2) language speakers (H. R. Pollio, 1977; H. R. Pollio & Smith, 1980). Arguably, not until Low's (1988) proposal of several metaphor-related skills (developed further in Littlemore & Low, 2006a, 2006b) did researchers began to seek out the potential of metaphor for second language learning. To date, L2 metaphoric competence has been operationalised in terms of fluid mental processes occurring when metaphors are comprehended and produced in speaking (e.g., Johnson & Rosano, 1993; Littlemore, 2001), and as the quality of interpretations and productions when test takers are given time and work in the written mode (Azuma, 2005; Hashemian & Nezhad, 2007; Zhao et al., 2014). L2 metaphoric competence research has focused mainly on English as the target language, and involved learners from a variety of L1 backgrounds (e.g., Japanese, French, Mandarin Chinese, and Persian).

Research has shown generally strong correlations between L2 receptive metaphoric competence and L2 proficiency (Zhao et al., 2014), and L2 productive metaphoric competence and L2 vocabulary *depth* (Azuma, 2005). However, the extent to which L2 vocabulary *size* and *depth* (see section 2.3.4), and L2 proficiency predict L2 metaphoric competence (as a combined model or separate predictors) remains unclear. In addition, correlations between L2 receptive and L2 productive metaphoric competence in intermediate learners have been both medium-to-large (Azuma, 2005) and negligible-to-small (Littlemore, 2001).[1] However, the extent to which the strength of relationship between the two modes changes from lower to higher L2 proficiency levels, and what this might suggest about the development of L2 metaphoric competence remains unknown.

Moreover, Low (1988) and Littlemore and Low's (2006a, 2006b) suggested metaphor-related skills and (sub)competences have existed for 29 and 11 years respectively, but have never been elicited empirically, or used to develop tests of L2 metaphoric competence. Rather, L2 metaphoric competence instruments have been very limited in their scope (e.g., Aleshtar & Dowlatabadi, 2014; Azuma, 2005; Hashemian & Nezhad, 2007; Zhao et al., 2014). Only one study to date (Kathpalia & Carmel, 2011) has measured Littlemore and Low's (2006a, 2006b) (sub)competences in the English writing of Singaporean university students. As a result, it is unclear whether Low (1988) and Littlemore and Low's (2006a) constructs can be reliably elicited

[1] For information on terms used to categorise correlation strength, see section 7.2.4.2.

and measured, to what extent the frameworks suggested are empirically supported, and whether L2 metaphoric competence is in fact underpinned by more basic fundamental (sub)competences that the authors were not in a position to detect.

Further issues in metaphoric competence research concern mixed levels of instrument reliability (e.g., Littlemore, 2001), non-reporting (e.g., Johnson & Rosano, 1993) and misreporting of instrument reliability estimates (e.g., Aleshtar & Dowlatabadi, 2014). Given that in quantitative research, determining the extent to which the data warrant and sustain a purported finding or explanation depends on (among other things) having reliable instrumentation (Plonsky & Derrick, 2016), these are significant shortcomings. Furthermore, tests of L1 and L2 metaphoric competence developed for particular studies (Azuma, 2005; Johnson & Rosano, 1993; Littlemore, 2001; H. R. Pollio & Smith, 1980) seem to require substantial refinements, in one case resulting in a 50% reduction of questions after piloting (Littlemore, 2001). Why are L2 metaphoric competence tests so hard to develop?

In response to these issues, the present study, using a sample of L1 Chinese non-native speakers (NNSs) of English (see section 1.2) and English native speakers (NSs), addresses the following research questions (RQs):

1) To what extent can (L1 and L2) metaphoric competence be reliably elicited and measured?
2) How do metaphoric competence test scores appear to differ between groups of English NNSs (L1 Chinese) and NSs of English?
3) To what extent do factors underlie the observed L2 metaphoric competence, vocabulary knowledge and proficiency test scores for the NNSs? What kind of (sub)competences might these factors represent?
4) To what extent can the same factors be found in the NNS and combined NNS+NS data, and how do the NNSs' and NSs' factor scores differ?
5) To what extent can L2 vocabulary knowledge (size and depth), L2 proficiency (Oxford Online Placement Test and IELTS), age of starting to learn English and time spent living in the UK predict L2 metaphoric competence test scores?
6) To what extent is the relationship between L2 receptive metaphoric competence and L2 productive metaphoric competence different at various L2 proficiency levels?

1.2 The educational context: International students in the UK

At the time of this research, the UK Council for International Student Affairs (2017) reported that of the students studying in Higher Education in the UK, 81% were home students, 6% were from EU countries, and 14% were from the rest of the world. For postgraduates, proportions were very different, with 46% of students coming from outside of the EU. At 91,215, the number of Chinese students far exceeded any other nationality, with China as the only country showing significant increases in student numbers. It is partly for these reasons that the NNS demographic studied in this thesis comprised L1 Chinese learners of English. This fact alone also means that the present study's findings are directly relevant to the largest demographic of international students in the UK. Although international students from non-English speaking countries are formally required to prove that their English is above CEFR (The Common European Framework of Reference for Languages) level B2 (2017), in order to receive a Tier 4 (General) student visa, acceptance on many university courses requires demonstrating higher levels of L2 proficiency than this.

At the University of York, where the research took place, the Centre for English Language Teaching (CELT) supports departments by providing pre- and in-sessional courses aimed at helping students integrate into the international environment of the university, become full members of the international academic community, and develop a global perspective and global skills in preparation for their future careers. Outside of their timetabled learning, international students have numerous opportunities to interact with other (home, EU and non-EU) students through events organised by the University of York Students Union and Graduate Students' Association, and the college joined upon enrolment. Many students also indulge in extensive travel within the UK, sightseeing and non-university activities (e.g., organised international cafes) where they interact with local and national residents.

Seemingly, these circumstances would offer international students ample opportunity for improving their L2 metaphoric competence. Although there is evidence to suggest that time spent living in an L2 immersion setting will positively impact on the diversity of lexis that learners produce, and help them become sensitised to nativelike word combinations (Foster, Bolibaugh, & Kotula, 2014; Foster & Tavakoli, 2009), there are indications that L2 metaphoric competence might not develop so easily (Azuma, 2005). The present study provides some new information on this complex issue.

1.3 Outline of the thesis

This chapter has established the research context and background information on studying as an international student in the UK. Chapter 2 will present a review and critique of the relevant

research literature, divided into four sections: (1) foundational issues on identifying and analysing metaphor in language, thought and communication; (2) an overview of L1 and L2 metaphoric competence research to date; (3) a synthesis of L2 vocabulary knowledge research; (4) a synthesis of L2 proficiency research. Chapter 3 presents a second literature review chapter, comprising a small study on instrument, interrater and intrarater reliability in L1 and L2 metaphoric competence research, leading up to an identification of the research gaps and formation of the research questions. Chapter 4 presents the methodology, outlining the rationales for various decisions, the development of the Metaphoric Competence (MC) Test Battery and the actual method used. Chapter 5 presents the results of Analysis 1, concerning the development and reliability of the MC Test Battery and descriptive statistics. These results answer the first two research questions, and a critical discussion of them follows in Chapter 6. Chapter 7 presents the results of Analysis 2, concerning metaphoric and other (sub)competences uncovered. These results answer the third and fourth research questions, with a critical discussion to follow in Chapter 8. Chapter 9 presents the results of Analysis 3, concerning the relationships between L2 metaphoric competence, vocabulary knowledge, general proficiency, age of starting to learn English and time spent living in the UK. These results answer research questions five and six, and are discussed critically in Chapter 10. Finally, Chapter 11 summarises the study and main findings, and presents limitations, further research needed, some tentative teaching implications and the main contribution of the study to the field of L2 metaphoric competence research.

Chapter 2: Literature Review 1

2.1 Foundational issues

2.1.1 Metaphor in language

Generally speaking, metaphor can be seen as the process of "treating X as if it were, in some ways, Y" (Low, 1988, p. 126). Metaphor in language (linguistic metaphor) is any word, phrase or utterance whose meaning appears to be incongruous in context, but is nonetheless understandable through some connection to the meaning of the surrounding discourse (Cameron, 2003). Cameron's (2003) example of a primary school maths teacher encouraging her students by saying "you're on the right track" (p. 3) provides an illustrative example of this. In this utterance, the discourse-appropriate interpretation (or what is actually meant), *you're working correctly*, and the fact that a "track" ordinarily has *trains* or *athletes* on it, makes it incongruous with the surrounding discourse. Linguistic metaphors are often said to comprise a *Topic* (the idea being expressed, not always mentioned explicitly) and a *Vehicle* (the actual language used to express the idea). Isolating metaphor in language is complicated by its co-occurrence with tropes such as simile (e.g., 'as good as gold'), metonymy (e.g., 'the White House issued a statement'), irony, sarcasm and hyperbole, and the fact that metaphor is a matter of degree, rather than a dichotomous, either-or phenomenon (Cameron, 2003; Carter & McCarthy, 2004; Littlemore & Low, 2006a). As shall be seen shortly, it is only by pinpointing the basic and contextual meanings of words, that one can reliably identify metaphor in language.

2.1.1.1 Identifying metaphor in language (linguistic metaphor)

Cameron's (2003) assertion that the basic or central meaning of the Vehicle "track" is "an athletics track" (p. 3) is problematic. One might well ask, *why is an athletics track the most basic meaning, and not a rough path or road*, *marks left by a person or animal*, *a metal construction that trains travel on*, *a song appearing on a record/CD*, or *a pole or rail that a curtain moves along?* It was this kind of dilemma, how the basic and contextual senses of words can be ascertained, that led to the development of more principled ways of identifying linguistic metaphor.

2.1.1.2 MIP, MIPVU and VIP: Three procedures for identifying linguistic metaphor

MIP

The product of several years' work, Metaphor Identification Procedure (MIP), developed by a

team of metaphor scholars known as the PRAGGLEJAZ[2] group (2007), and its later refinement MIPVU[3] (2010), developed by Gerard Steen and a team of PhD students working at the Vrije Universiteit (Amsterdam), offer reliable methods for identifying language that is 'structurally' metaphorical (in terms of contextual and basic senses), and for comparing the frequency of linguistic metaphor across various texts and discourses. MIP is operated as follows (Pragglejaz, 2007, p. 3):

1. Read the entire text–discourse to establish a general understanding of the meaning.
2. Determine the lexical units in the text–discourse.
3. (a) For each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.
   (b) For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be
   —More concrete (what they evoke is easier to imagine, see, hear, feel, smell, and taste);
   —Related to bodily action;
   —More precise (as opposed to vague);
   —Historically older;
   Basic meanings are not necessarily the most frequent meanings of the lexical unit.
   (c) If the lexical unit has a more basic current–contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.
4. If yes, mark the lexical unit as metaphorical.

The authors provided a worked example in which 28 lexical units[4] (separated by '/' below) are identified in the first sentence of a British newspaper article. Six of these lexical units (underlined) were classified as metaphorically used (pp. 3-13):

> / For / years, / Sonia Gandhi / has / struggled / to / convince / Indians / that / she / is / fit / to / wear / the / mantle / of / the / political / dynasty / into / which / she / married, / let alone / to / become / premier.

The authors acknowledged several limitations of MIP. First, as a procedure requiring a binary decision - *yes this a metaphor, or no this is not* - MIP runs contrary to the prevailing view of metaphor as a matter of degree (Littlemore & Low, 2006a), and in this respect is overly reductive. Second, although it conveys the sense that writers' or speakers' linguistic metaphors

[2] Named according to the initial letters of the team: Peter Crisp (Chinese University of Hong Kong), Raymond Gibbs (University of California, Santa Cruz), Alice Deignan, (University of Leeds), Graham Low (University of York), Gerard Steen (Vrije University of Amsterdam), Lynne Cameron (University of Leeds/The Open University), Elena Semino (Lancaster University), Joe Grady (Cultural Logics), Alan Cienki (Emory University), and Zoltán Kövecses (Loránd Eötvös University).

[3] The additional 'VU' refers to the institution where the procedure was developed: the Vrije Universiteit (Amsterdam).

[4] In the published article, the number of lexical units is erroneously reported as "27" (p. 13).

have been discovered, MIP in fact makes no claim about the extent to which writers or speakers intended their specific words to express metaphorical meanings. Third, the decision not to mark a word as metaphorical does not imply that it has been used literally, since it may be expressing a metonymic, hyperbolic or other figurative meaning, which MIP does not concern (Pragglejaz, 2007).

MIPVU

In response to the third of these limitations in particular, MIPVU was developed. In MIPVU, metaphor-related words (MRWs) that are lexical expressions of underlying cross-domain mappings, are identified. These include (Steen et al., 2010, p. 25):

1. Indirect metaphor (e.g., A is B type metaphor), essentially metaphors that would be identified by MIP
2. Direct metaphor (e.g., simile) whereby a word's use may potentially be explained by some form of cross domain mapping to a more basic referent or topic in the text.
3. Implicit metaphor whereby a word used for the purpose of lexico-grammatical substitutions (e.g., a third person pronoun) is also a direct or indirect metaphor
4. Metaphor flags (Mflags) which are lexical units such as tuning devices that signal that a cross domain mapping may be at play

MIPVU uses a similar series of steps to place lexical units in these categories, or identify them as not metaphorically used. Importantly, MIPVU does not deal with conventionality or creativity, it merely identifies whether a lexical unit is, or is not, an MRW. In acknowledgement of the inherent ambiguity in metaphor identification, and to deal with problems such as broken utterances, the authors provided a further category, *when in doubt, leave it in* (WIDLII). Subsequent authors using MIPVU (e.g., Nacey, 2013) have praised its developers' recommendation that analysists hold meetings to discuss decisions and problem cases, and have vouched for the usefulness of the online discussion forum and bank of problem cases created by the team.

MIPVU has led to high reliability between and within raters, both in its original application to academic, news, literary and spoken discourse genres and in further applications (e.g., Nacey, 2013). However, as a time-consuming, tedious procedure not yet capable of being performed by a computer, MIPVU is subject to (at least) two main criticisms. First, because of the time involved in coding, analysts are rarely able to have all of their data recoded by a second or third rater, or recode it themselves. Second, as the MIPVU authors acknowledged, because MIPVU does not permit the identification of metaphor at the morphological level, potential metaphors within words (e.g., the prefix 'over-' in 'overstatement') go unidentified (Steen et al., 2010, p. 189).

VIP

Although MIP and MIPVU have brought statistical rigour to the field of metaphor identification, they are often felt to be too restrictive. Several years prior to the emergence of these procedures, Cameron's (2003) Vehicle Identification Procedure (VIP) had been used to identify metaphor in the discourse of British primary school children and their teachers. VIP requires the coder to read through the discourse, underlining possible Vehicle terms according to a set of malleable rules on what will count as metaphor. Unlike MIP and MIPVU, VIP does not require the text to be divided into lexical units, and so there is no restriction on the word limit of a Vehicle term. In some cases, whole utterances, sentences or paragraphs can be underlined as metaphorical. Whereas MIP and MIPVU result in a percentage of metaphorically used lexical units, with VIP, frequencies tend to be reported as X amount of Vehicle terms underlined per 1,000 words. Limitations fundamentally relate to VIP's lack of rigour and that Vehicle term frequencies are not comparable across different studies. VIP is much less commonly used than MIP or MIPVU, but has been applied by researchers other than Cameron, for instance, to identify clusters of metaphors in Baptist sermons (Corts & Meyers, 2002).

2.1.2 Metaphor and thought

2.1.2.1 Conceptual metaphor theory

Until only a few decades ago, metaphor was primarily regarded as a poetic and ornamental (but ultimately superfluous) feature of language. One of the lasting achievements of Lakoff and Johnson's (1980) *Metaphors We Live By* was to highlight the pervasiveness of metaphor in human language and its fundamental role in human cognition. Nacey (2013) notes that although the authors were not the first to realise that metaphor plays a role in thought, their finding that a multitude of metaphors in language could be theoretically grouped into a relatively small number of conceptual metaphors (probably several hundred according to Gibbs, 2011) was a novel one. Conceptual metaphor theory (CMT), as this came to be known, purports that metaphorical conceptualisation involves the use of an entity from one source domain used to understand an entity from a semantically unrelated target domain. Thus, 'I need to spend more time writing my thesis' is indicative of the target domain TIME being understood via a conceptual mapping from the source domain MONEY, and the conceptual metaphor TIME IS MONEY[5]. Metonymy, in the CMT view, occurs when the source and target domains are identical or highly overlapping. For instance, 'nice wheels' involves one aspect from the domain of CARS ('wheels') standing for the wider domain.

CMT has been remarkably successful at accounting for conventional expressions in

[5] Conceptual metaphors and their individual component concepts are conventionally capitalised.

multiple languages, and seems to have distinct potential for L2 language learning. Boers (2000), for instance, found that a group of L1 Dutch intermediate learners of L2 English informed about the conceptual metaphors THE BODY IS A CONTAINER FOR THE EMOTIONS and ANGER IS A HOT FLUID IN THE CONTAINER performed better at recalling single and multi-word metaphors such as "she erupted" and "she flipped her lid" (p. 563) on a cloze-test than a control group. Boers (2000) also found that a group of L1 French intermediate learners of English informed about conceptual metaphors for financial reporting (e.g., PROFITS ARE AIRCRAFTS) produced more target linguistic metaphors such as "soar" and "skyrocket" (p. 558) in written essays than a control group, although both groups were on a par in terms of inaccurate productions. In a third experiment, a group of L1 French intermediate learners of English who were given a set of prepositional and phrasal verbs presented under the headings of orientational conceptual metaphors (e.g., MORE IS UP, LESS IS DOWN, ACTIVE IS UP, INACTIVE IS DOWN) were better at supplying the intact verbs from their notes in a subsequent cloze-test than a control group.

In spite of experimental successes such as these, CMT has been attacked almost since its inception. Criticism concerns its inability to explain language in real use, the tendency of its proponents to use fabricated examples to prove its points, and its imposition of conceptual metaphors on linguistic data in an unsupported way (Cameron & Deignan, 2006; Gibbs, 2011; Littlemore & Low, 2006a). Others have argued that CMT is essentially circular and its claims impossible to falsify (Murphy, 1996; Vervaeke & Kennedy, 1996). To the extent that CMT deems the actual words of linguistic metaphors as unimportant, it is further problematised by grammar. For instance, from the Bank of English corpus, Littlemore and Low (2006a, p. 174) observed that the verb 'leak' adopts collocational structures of 'communicate', a meaning it metaphorically conveys (e.g., 'Washington leaked the fact that', 'a widely leaked email', 'the news was leaked by employees', 'when word leaked out that').

2.1.2.2 How is metaphor processed?

Several theories have been developed to explain how metaphor is processed. These include the literal first model which states that a figurative interpretation will only come into play if a literal interpretation is shown to be false (Searle, 1993); the direct access model which posits that figurative senses can be activated before literal ones in certain contexts; the graded salience hypothesis which predicts that the speed of access depends on the salience of a meaning in the speaker's mind (Giora, 1999, 2003); the comparativist view which holds that metaphors (as condensed forms of comparison) are equivalent in meaning to their counterpart similes (Glucksberg & Haught, 2006); the career of metaphor theory which predicts that a community of language users will first process a new metaphor as a comparison and move over time to processing it as a categorisation as the metaphor becomes more conventionalised (Bowdle &

Gentner, 2005); and the quality-of-metaphor hypothesis which predicts that the most apt metaphors work best (and in some cases only) as categorisation assertions whereas poor metaphors work best as comparisons (Glucksberg & Haught, 2006).

Each theory has evidence for and against it. For instance, the direct access model is supported by findings indicating that native speakers encountering unconventional uses of idioms tend to analyse the idiomatic meaning of these expressions before deriving the literal, unconventional interpretation (e.g., Gibbs, 1980); the literal first model is challenged by findings that the truthfulness of L1 statements appears to be no more quickly recognised via literal than figurative language (McElree & Nordlie, 1999); and the career of metaphor theory is undermined by experiments which show that novelty, per se, does not privilege comparison over categorization, suggesting instead a role for aptness and ease of comprehension (Glucksberg & Haught, 2006). Ultimately, while most theories can be roughly categorised in terms of whether or not they suggest that the literal meanings of metaphors first need to be rejected, there appears to be little field-wide consensus on how metaphors are processed (Littlemore & Low, 2006a).

2.1.3 Metaphor in communication

2.1.3.1 Deliberate metaphor theory

Around a decade ago, Steen (2008) proposed a three-way model of language, thought and communication, and deliberate metaphor theory (DMT). DMT suggests that *deliberate* metaphors are linguistic metaphors that "explicitly invite the addressee to conceptualise one thing as another thing, often for rhetorical or persuasive purposes" (Steen, 2008, p. 213), whereas non-deliberate metaphors do not. Deliberate metaphors can either be *novel* or *conventional* (discussed below), and in contrast to *non-deliberate* metaphors, are claimed to be processed as online cross-domain mappings. Since analysts cannot directly know the mind of the speaker, deliberate metaphors must be identified via linguistic clues such as the use of analogy, simile, tuning devices, novel metaphor, metaphor clustering or repetition. A formal procedure for the identification of deliberate metaphor, incorporating some of these clues, is currently pending publication (Reijnierse, Burgers, Krennmayr, & Steen, under review).

Over the past decade or so, DMT has been refined and advanced in a series of publications as a useful tool for exploring new ideas about metaphor and re-interpreting the findings of existing research (Steen, 2008, 2011a, 2011b, 2013, 2015, 2016, 2017). However, the theory has been criticised for ignoring or not being able to account for a vast body of cognitive linguistic and cognitive science empirical findings, not paying due attention to communication and consciousness, and seeking to re-orient metaphor as an ornamental and poetic feature of language, produced only by specialists (Gibbs & Chen, 2016). DMT's proponents have

responded, in turn, that many of its criticisms are based on misrepresentations or misunderstandings of its basic tenets, and that the theory does not (as is claimed) seek to fundamentally separate metaphor in language, thought and communication (Steen, 2017).

2.1.3.2 Historical, entrenched and novel metaphors

A final foundational issue concerns the distinction that some authors have made between *historical* (or *dead*) metaphors, *entrenched* metaphors and *novel* metaphors (Lakoff & Turner, 1989; Müller, 2008). Historical metaphors are semantically opaque words or expressions which once had a (now defunct) literal sense, from which a metaphorical sense emerged, and which remains the only meaning in contemporary language usage (e.g., 'to show someone the ropes', an expression originally related to teaching someone to master the rigging on ships). From a structural (i.e., MIPVU) perspective, since only their figurative sense remains, historical metaphors are not metaphors in contemporary usage (Nacey, 2013). Entrenched metaphors are usually semantically transparent words or phrases, whose metaphorical and literal senses are codified in standard dictionaries. Both historical and entrenched metaphors can be regarded as conventional. Novel metaphors on the other hand, which also tend to be transparent, do not have their contextual senses codified in dictionaries.

2.2 Metaphoric competence (L1 and L2)

2.2.1 What is metaphoric competence?

Perhaps the most common way of defining *metaphoric competence* (sometimes called *metaphorical competence*) is as the ability in the L1, L2, L3[6] or otherwise, to comprehend and produce metaphors in language, thought or communication (Littlemore & Low, 2006a). By implication, metonymic, idiomatic, and figurative language competences can be defined as the abilities to comprehend and produce these tropes (H. R. Pollio & Smith, 1980). A second and more nuanced conceptualisation of metaphoric competence, and the one that will be used in the present study, is as a set of metaphor-related skills and (sub)competences. This view was advanced in Low's (1988) article "On teaching metaphor" and Littlemore and Low's (2006a) book *Figurative Thinking and Foreign Language Learning* (and related article, 2006b). These publications drew attention to the semantic, syntactic and pragmatic behaviour of metaphor in the language of native and non-native speakers, and set out theoretical frameworks of metaphoric competence intended to be useful for second language learners, teachers, and researchers alike. Although neither study was empirical in itself, the characterisations of metaphoric competence provided were based on a broad range of findings from the available

[6] The 'L3' refers to another (less well-known) second or foreign language known by a learner.

research literature.

Because of their importance to L2 metaphoric competence research and the present study, Low (1988) and Littlemore and Low's (2006a, 2006b) studies are first critically reviewed. Following that, I synthesise empirical studies on L2 metaphoric competence in relation to second language proficiency, vocabulary knowledge, conceptual and phraseological fluency, cognitive style and individual traits, studied in both experimental and naturalistic settings.

2.2.2 Two influential theoretical accounts of L2 metaphoric competence

2.2.2.1 "On teaching metaphor" (Low, 1988)

In his 1988 article, Low argued that metaphor should be given a more prominent role in language teaching because of its centrality to language use, the fact that it pervades large parts of the language system, and enough (by the time of writing) had been discovered about it to make this possible. The author began by offering a working definition of metaphor and describing some of its functions (e.g., to make it possible to talk about abstractions, to allow the speaker to discuss emotionally charged subjects). Low then hypothesised several metaphor-related skills that seem to characterise L2 metaphoric competence (pp. 129-135):

1. Ability to construct plausible meanings
2. Knowledge of the boundaries of conventional metaphor, involving:
    a) Knowledge of which feature of the Vehicle Y can be exploited conventionally
    b) Knowledge of Vehicles used to describe more than one Topic
    c) Knowledge of Vehicle acceptability across different word classes
    d) Knowledge of mixed metaphors
3. Awareness of acceptable Topic Vehicle combinations
4. Ability to interpret and control hedges
5. Awareness of socially sensitive metaphors
6. Awareness of multiple layering in metaphors and interactive awareness of metaphor.
7. Interactive awareness of metaphor

For the teaching of conventional metaphor, Low concluded by advocating consideration of structural (e.g., grammatical) aspects, awareness of boundaries (i.e., what is not normally said), and reasons why certain metaphors do not seem to mix well. He also called for more research into how native speakers react to novelty and innovation in metaphor.

Despite being the first significant discussion of the potential of metaphor for the English as a foreign language (EFL) classroom, Low's study can be criticised on several grounds. First, empirical evidence is required to verify his claims that "native speakers are frequently expected to be good at…" or "learners [to some degree] need to develop" (p. 129) the skills he described,

that mixing metaphors frequently "presents a problem for both native speakers and language learners" (p. 131). Empirical evidence is also needed to assess the extent to which several of his linguistic examples are in fact "acceptable utterances" (p. 130). Second, his arguments that L2 learners need to acquire "knowledge of the boundaries of conventional metaphor - what people tend not to say" (p. 130), and that "learning 'one-off' examples does not help learners resolve the structural problem of where the boundaries of a metaphor are felt to lie, nor how rigid native speakers perceive particular boundaries as being" (p. 137-138) carry the assumptions that such boundaries exist, and that native speakers have a shared knowledge of them. This also needs investigating.

Another problem (although not a criticism of Low's arguments per se) concerns the extent to which it is useful to compare L2 metaphoric competence with the language of native speakers. One the one hand, Low's examples of Chinese and English as "first and target" (p. 136) languages, and his discussion of the ways in which native speakers comprehend and produce metaphors imply that he is advocating that L2 learners should become familiar with (and perhaps emulate) native speaker norms. On the other hand, his definition of *competence* in terms of being an accepted, interesting member of one's social groups, and his acknowledgement that L1 transfer may be recognised as conscious innovation (rather than second language error, which it may also be) suggest that L2 learners need not necessarily aspire to native speaker norms (see section 2.4.4).

2.2.2.2 Figurative Thinking and Foreign Language Learning (Littlemore & Low, 2006a)

In their 2006 book and related journal article, Littlemore and Low argued that metaphor is relevant to all four components of Bachman's (1990) model of Language Competence[7], involving grammatical, textual, illocutionary, and sociolinguistic (sub)competences (Table 2.1):

Table 2.1 *Bachman's Model of Language Competence (1990, p. 87)*

Organizational competence		Pragmatic competence	
Grammatical competence	Textual competence	Illocutionary competence	Sociolinguistic competence
Vocabulary or variety	Cohesion	Ideational functions	Sensitivity to dialect
Morphology	Rhetorical organization	Manipulative functions	Sensitivity to register
Syntax		Heuristic functions	Sensitivity to naturalness
Phonology/ graphology		Imaginative functions	Ability to interpret cultural references and figures of speech

[7] Bachman (1990, p. 87) proposed that Communicative Competence = Language Competence + Strategic Competence.

Bachman specifically located metaphor within sociolinguistic competence, as part of the ability to interpret cultural references and figures of speech. Concerning illocutionary competence (one's ability to understand the message being communicated), Littlemore and Low argued that metaphor plays a role in the ability to use metaphor to communicate an emotional standpoint (ideational functions), the use of trial and error and ad hoc devices to learn and teach others about the world around us (heuristic functions), and the ability to use language to create an imaginary world or extend the world around us for reasons of humour or aesthetics (imaginative functions). Lexico-grammatical competence,[8] the authors argued, involves metaphorical processes in demonstratives, prepositions and particles, phrasal and prepositional verbs, tense and aspect, modality, and phraseological patterning. Metaphor is located within textual competence as part of cohesion and rhetorical organisation. Finally, using a more general conceptualisation of strategic competence than Bachman (1990), the authors point to the role of metaphor in linguistic compensation strategies (e.g., word coinage, circumlocution, and transfer from the L1).

Unfortunately, the authors gave no quantitative indication on the centrality and relative importance of metaphor to each of these components, or of the relative importance of the components themselves to language competence, something that Bachman was also criticised for (Skehan, 1998). In this respect, it is also unclear which (sub)competence, if any, might reveal the greatest metaphor-related differences between L1 and L2 speakers, or the extent of conceptual overlap between the various framework components.

2.2.3 Research into L2 metaphoric competence

Metaphoric competence research can be defined as any study that investigates first or second language learners' awareness, retention, comprehension and/or production of metaphor and other figurative language. Although authors often use the specific term *metaphoric competence* to refer to their research as such, in many studies this is not the case. Under this definition, metaphoric competence studies are roughly distinguishable into two types. Those which use *elicitation* methods involve the use of tests and experimental stimuli to gather and measure understandings or productions of specific metaphors. While elicitation methods may be criticised for decontextualising metaphoric competence from real world language use, they allow for the targeting of specific metaphors and aspects of metaphoric competence. *Naturalistic* methods, which explore metaphoric competence *in the wild*[9] by measuring patterns of metaphor use in free, unprompted spoken or written production, provide a better indication

[8] Littlemore and Low (2006a, 2006b) used the term 'lexico-grammatical', whereas Bachman used 'grammatical'.

[9] This expression is used deliberately to draw a parallel with Steen et al.'s (2010) discussion of MIPVU and metaphor in spoken discourse.

of real-world language use, but offer no guarantees that the speaker (or corpus) will yield instances of the metaphors or skills of interest. This distinction is not perfect (e.g., an essay title might be argued to *elicit* free, *naturalistic* production), but it is useful for highlighting an important methodological choice facing researchers.

2.2.3.1 …and L2 vocabulary knowledge

Apparently, the only study to specifically compare elicited L2 metaphoric competence with L2 vocabulary size and depth is Azuma's (2005) investigation into the metaphoric competence[10] of 42 (pilot study), 57, 56 and 59 (main study) Japanese EFL learners. In designing her metaphoric competence tests, Azuma summarised, but eventually rejected, Low's (1988) metaphor-related skills and Littlemore's (2001) metaphoric competence tests, instead using tests in her study from "two experiments which greatly inspired me" (p. 112). Metaphoric competence was measured via Metaphorical Competence Receptive and Productive Tests (MC-RT and MC-PT) (based on Gibbs, 1980) measuring ability to explain the meaning of literal and figurative senses of idioms in two passages (MC-RT) and write two passages embedding idioms conveying these senses (MC-PT). Both tests were scored 0-3 using specially developed partial-credit criteria. Metaphorical competence was also measured using the MC-XY Test (based on Winner, Rosentiel, & Gardner, 1976) measuring ability to write two sentences (one using an adjective literally, one figuratively) in the format *X is a(n) adjective Y*. Vocabulary size (see section 2.3.4) was measured via the Vocabulary Level's Test (VLT) (Schmitt, 2000; Schmitt, Schmitt, & Clapham, 2001). Vocabulary depth was measured using a simplified version of the Word Associates Test (WAT) (Read, 1993), and the Polysemy test (PolyT) (developed by Azuma), measuring ability to translate 10 WAT words (6 adjectives, 2 verbs and 2 nouns) from Japanese to English and describe the "state, act or situation meant by the word or the sentence" (p. 111).

The study produced several key findings. Participants were much better at understanding metaphor than producing it. A combined MC-RT and MC-PT variable correlated most strongly with vocabulary depth measured by the PolyT (r = .67, .82, .52) and vocabulary size (r = .58, .78, .50), but less so with the other metaphoric competence measure (MC-XYT) (r = .53, .53, .38) and vocabulary depth measured by the WAT (r = .28). Participants found the simplified version of the WAT anxiety inducing, leading to its removal before the main study. MC-RT was slightly more strongly associated with vocabulary depth (PolyT) (r = .58, .74, .42) than vocabulary size (the VLT) (r = .48, .72, .42). The MC-PT was also more strongly correlated with vocabulary depth (r = .42, .73, .31) than vocabulary size (the VLT) (r = .39, .60, .28). Correlations between the MC-XYT and the vocabulary measures varied (ranging from r = .33 to .52) but groups were not homogenous in terms of which vocabulary knowledge measure had the

[10] Azuma (2005) used the term *metaphorical competence*.

strongest correlation with this MC measure. Finally, the MC-RT and MC-PT had what might be considered as medium, strong and medium correlations (see section 7.2.4.2), significant at the .05, .01 and .01 levels for the three subgroups respectively (*r* = .33, .56, .37). In sum, these correlations suggest a fairly strong relationship between receptive and productive metaphoric competence for these learners.

In order to investigate the possible effect of living in an L2 immersion setting, Azuma quantitatively and qualitatively analysed data from five participants who had studied abroad. While acknowledging that this was an insufficient number of participants to draw hard and fast conclusions from, the author nonetheless suggested that the period that these participants spent studying abroad (less than one year) did not sufficiently impact any aspect of their L2 metaphoric competence.

The strengths of Azuma's study lie in the detail of reporting, the fact that tests were refined and administered to more than one sample of participants (allowing for replication of findings), and the abundance of data provided including test scores, examples of response patterns, participants' and teachers' attitudes towards the tests for pedagogical and validation purposes. However, the general scope of the study, particularly in terms of its measurement of L2 metaphoric competence, is limited in several ways. First, although Azuma argued that all target items were representative of metaphors that Japanese EFL learners are likely to encounter when using dictionaries as learning materials, several items appear to be specific to English literature of the past few hundred years rather than contemporary usage, for instance "the rotten apple injures its neighbours" (pp. 338-342). Second, some items are used with peculiar senses, for instance "a little pot is soon hot" (pp. 338-342) to mean *people (especially girls) soon grow up and become attractive* rather than *a small person is easily roused to anger or passion* which Oxford Reference Dictionary (2017) suggests. Third, some of the items presented in the MC-RT are unusually formed, for example, "you cannot eat your cake and have your cake" (p. 341) rather than 'you cannot have your cake and eat it (too)', suggested by Macmillan English Dictionary (MED) .

2.2.3.2 …and L2 proficiency

Zhao, Yu and Yang (2014) conducted research into the relationship between the L2 receptive metaphoric competence and L2 reading proficiency of 75 L1 Chinese learners of English. While the authors acknowledged Low's (1988) metaphor-related skills as "a pioneering examination of metaphorical competence" (p. 169), L2 receptive metaphoric competence was measured simplistically, using Azuma's (2005) MC-RT and MC-XYT (described above). L2 reading ability was measured using the reading section of a language test developed by English teachers and professors in their institution involving cloze and comprehension questions.

The researchers found a positive correlation between the MC-RT and the reading test (*r* = .43), significant at the .01 level, leading to their (somewhat vague) conclusion that L2 receptive metaphoric competence is linked to L2 proficiency. A non-significant, negligible correlation between the MC-XYT and reading comprehension test was also observed (*r* = .01), which was interpreted as reflecting the fact that the MC-XYT in a sense, involves the production of language and thus would not have been expected (by the authors) to correlate with L2 reading comprehension.

In another study, Aleshtar and Dowlatabadi (2014) measured the relationship between the L2 metaphoric competence and L2 proficiency of 60 L1 Persian undergraduate learners of English. The authors acknowledged several perspectives on L2 metaphoric competence, but erroneously interpreted Low (1988) and Littlemore and Low (2006a, 2006b) as arguing for metaphoric competence "as the third competence [after grammatical and communicative]" (p. 1897) rather than as pervading large parts of the language system (Low, 1988) or playing an important role in all areas of communicative competence (Littlemore & Low, 2006a, 2006b).

L2 metaphoric competence was measured by NourMohamadi's (2010) English Conventional Metaphor Proficiency Test (ECMPT), an instrument consisting of six sections (each containing 15 items) relating to six types of variation reported by Kövecses (2003) to exist between metaphors in two languages. Problematically, the authors did not give any indication as to whether this test measured metaphoric competence in the written or spoken modes, and reported the reliability of the instrument when used in NourMohamadi's (2010) study rather than when used in their own. The L2 proficiency measure used was a 2001 version of the Oxford Online Placement Test (OOPT, see section 2.4.2.3).

The results showed large correlations between L2 metaphoric competence (as measured) and the L2 proficiency of supposedly "low" and "high" (p. 1898) proficiency groups (*r* = .77, and .72 respectively). In fact, the authors did not really discuss the significance of the correlations observed for the different proficiency groups. However, the fact that these groups were formed arbitrarily, and that the OOPT scores for both overlapped, severely limits any potential for such discussion.

2.2.3.3 …and cognitive style

Using L1 French learners of English, Littlemore (2001) conducted research into the relationship between both L2 and L1 metaphoric competence, cognitive style and communicative language ability. The author focused on the "fluid mental processes involved in metaphor production and comprehension…[thus using a] definition…narrower than that proposed by Low (1988), who includes aspects of crystallised intelligence" (p. 461). Thus, in contrast to other authors, Littlemore was explicit about why Low's (1988) metaphor-related skills were not used for test

development. Drawing on L1 metaphoric competence factors identified by Pollio and Smith (1980), the author defined metaphoric competence as the ability to think up one's own unconventional metaphors (Originality of Metaphor Production), find more than one meaning for a single given metaphor (Fluency of Metaphor Interpretation), think up possible meanings for novel metaphors (Ability to Find Meaning in Metaphor) and think up plausible meanings at speed and under pressure (Speed in Finding Meaning in Metaphor).

For the L1 metaphoric competence tests, instrument reliability ranged from Cronbach's alpha (α) = .58 (Originality of Metaphor Production) to α = .84 (Speed in Finding Meaning in Metaphor). For the L2 metaphoric competence tests, the range was α = .31 (Fluency of Interpretation) to α = .90 (Ability to Find Meaning in Metaphor).

The results showed that Littlemore's hypothesis that L2 metaphoric competence would be related to L2 communicative language ability was not supported. While some aspects of L1 metaphoric competence were related, this was not generally the case in the L2, for which only a small size correlation between Originality of Metaphor Production and Speed in Finding Meaning scores, significant at the .05 level, was found. The fact that Speed in Finding Meaning and Ability to Find Meaning in Metaphor correlated in the L1, but not the L2, was attributed to the retrieval process being less automatic in the second language compared with the first. The finding that participants with a holistic cognitive style were quicker at finding meaning in metaphor in both the L1 and the L2 was explained by the fact that Speed in Finding Meaning in Metaphor was the only metaphor test that was timed, and thus most comparable with the cognitive style tests. Finally, in line with the results of Pollio and Smith's (1980) Principal Components Analysis, Fluency of Metaphor Interpretation was completely independent from the other traits. The author explained that this finding was either due to this test being the only metaphoric competence measure that did not involve quality of interpretations, or because the true relationship with other variables was masked by the test's low instrument reliability.

2.2.3.4 …and L2 writing

A useful alternative to elicitation methods in L2 metaphoric competence research has been through the identification and analysis of metaphor in naturalistic data (e.g., spoken conversations, written assignments). The fact that Steen et al. (2010) found academic texts to be the richest in terms of metaphor frequency makes this genre particularly well-suited to metaphor research. In a study comparing metaphor production in academic A-Level assignments written by British English native speakers and L1 Norwegian learners of English, Nacey (2013) used MIPVU to analyse 20,243 words of text in the Louvain Corpus of Native English Essays (LOCNESS) and 20,468 in the Norwegian subset of the International Corpus of Learner English (NICLE). After the removal of 'for' and 'of' from the data (Steen et al., 2010), 13.3% of the words

in LOCNESS and 15.5% of the words in NICLE were coded as metaphorically related words (MRWs). Thus, there were more linguistic metaphors in the L2 English than the L1 English, suggesting that metaphor production is an important linguistic feature in the writing of all language users, both native and non-native. Prepositions were the most frequent word class of metaphor, with these, nouns and verbs comprising most of the metaphors in the scripts. One limitation of the study is that the patterns of metaphor use observed may have been influenced by students directly quoting other sources (e.g., learning materials), although the author argued this was not likely to have greatly impacted on the observed findings.

In a comparable study, Kathpalia and Heah (2011) analysed metaphor use in 113 samples of text written by Singaporean English as a Lingua Franca speakers on the topic "if you had a minute before an international audience, how would you prove yourself to be a worthy ambassador of the University?" (p. 275). The authors identified linguistic metaphor in relation to the four (sub)competences discussed by Littlemore and Low (2006a, 2006b), making this the only study to date to measure L2 metaphoric competence in the Bachman framework sense. Problems with grammatical/linguistic competence were the most common out of the four types studied, present in 99 out of 113 (88%) of the writing scripts. Illocutionary competence problems occurred in 95 (84%) of the scripts, textual competence problems in 70 (62%), and sociolinguistic competence in 21 (19%).

Overall, the study showed that although the learners made numerous attempts to use metaphor in their writing and to cover gaps in their English proficiency, on many occasions they appear to have lacked appropriate pre-fabricated language for doing so. Despite using "standard English" (p. 278) as criteria for determining target and interim forms, the authors were careful to point out that many of the miscollocations produced would be acceptable as Southeast Asian lingua franca forms of English, and that they do not intend to discourage their use in local or regional contexts. In conclusion, they suggest that mutual intelligibility, rather than nativelike proficiency, should drive the metaphors and idioms produced in more global (cross-linguistic) contexts. Problems with the study concern a lack of information about the actual procedure used to identify and reject metaphors, the number of raters and the extent to which they agreed, and questionable coding, for instance "…my fellow friends" classified as an interlanguage phrase but "…my fellow mates" (p. 279) as a target phrase.

2.2.3.5 …and conceptual fluency vs phraseological proficiency

Danesi (1992, 1995) proposed the notion of *conceptual fluency* to denote knowledge of how the target language encodes concepts on the basis of metaphoric reasoning. Development of this ability, he argued, can help overcome the "textbook literalness" (1995, p. 4) of learner language and mitigate conceptual inappropriateness stemming from unconscious transfer of native

language patterns. Critiquing this account, Philip (2010) argued that second language learners' problems with metaphor concern both lexical and conceptual aspects, thus emphasising the need to develop knowledge of the phraseological properties of metaphor, rather than concepts alone. In this view, the greater a student's repertoire of conventional collocations and phraseology, the more proficient they will be at expressing concepts effectively (section 2.4.3).

Using a longitudinal approach, Hashemian and Nezhad (2007) studied the development of conceptual fluency and metaphoric competence in L1 Persian learners of English. One group of 139 Junior students (presumably 16-17 years of age, although not stated explicitly) were found to have an increased conceptual fluency and higher L2 metaphoric competence test scores after attending classes for six months to learn about conceptual metaphor. These participants also had a higher average metaphor density in a paragraph written after the treatment than a paragraph written before. Moreover, the metaphor density of the latter paragraph matched that of a group of 23 English native speakers. Limitations of the study include the fact that no control group was used for comparison, little information was provided about the conceptual fluency and metaphoric competence test apart from that the latter involved receptive and productive sections and was a "teacher-made test comprising metaphors, idioms and the like" (p. 45).

In another study, Johnson and Rosano (1993) investigated the relationship between cognitive style involved in metaphor interpretation and second language proficiency, finding that L2 English participants (mixed L1s, but mostly Mandarin) performed less well to English native speakers on decontextualised oral measures of vocabulary and verbal analogies, but just as well at oral fluency and complexity of interpretation on a metaphor task. The results suggest that L2 proficiency in English (or L2 status in itself) should not always be seen as implying a deficit in L2 metaphoric competence.

2.2.3.6 …and language play

Language play has been described as repetition and manipulation of forms, semantic and pragmatic play, and banter and joking *in* (rather than simply *with*) the L2 (Cook, 1997, 2000). Language play can involve words, phrases, sentences, parts of words, groups of sounds, and series of letters (Crystal, 1998), and is also known to occur across both registers (Wray, 2008) and languages (Wang & Hyun, 2009).

Language play appears to be generally beneficial for second language learning. Research has found that engaging in it destabilises the learner's lexicon, thus allowing for restructuring of the interlanguage system (Bell, 2005, 2012; Kim & Kellogg, 2007; Tin, 2011) and helping prevent non-targetlike forms become fossilised (Selinker, 1972). Language play can also be used as a method to raise awareness of the relationship between L2 form and meaning (Sullivan, 2000)

and to this end, encourages both *noticing* of the existence of certain forms and *conscious awareness* of a rule or generalisation (Schmidt, 1994). Both processes can be exploited by teachers to help foster learning (Leow, 2000; Norris & Ortega, 2006). In addition, engaging in language play, especially in a humorous context, may result in deeper processing of lexical items, as the learner pays more and higher quality attention to forms, thus making them more memorable (Bell, 2005; Craik & Lockhart, 1972).

Another important aspect to language play is the role it serves in performing social functions and shaping L2 identity. Concerning social functions, engaging in language play can facilitate relationship building, where language is used to effect membership of social groups and conduct social actions within these groups (Belz & Reinhardt, 2004; Wray, 2012). Regarding L2 identity, both the intentional and incidental development of one's capacity to use language play to make jokes, express and understand opinions and feelings, and perform various other pragmatic functions in the L2 is inextricably linked to the development of the L2 personality (Bell, 2005).

Examples of language play: Idiom extension (Littlemore & Low, 2006a), a case in point

One does not have to look far to find examples of native speakers engaging in language play. For instance, Littlemore and Low's (2006a) observation that native speakers frequently play with language by referring to and extending the literal sense of idioms (and other linguistic metaphors) can be seen in an interview exchange on BBC Newsnight (2013) between the then presenter Jeremy Paxman and (now former) Mayor of London Boris Johnson, presented here with (likely) metaphorical language underlined:

> Paxman: …you said you would think about the Prime Ministership <u>if the ball came loose from the scrum.</u> Are you still <u>bound in the scrum?</u>
> Johnson: <u>The ball! The ball!</u> Shall I tell you <u>where the ball is now?</u>
> Paxman: Yeah do, and tell us <u>what position you're playing too!</u>
> Johnson: I'm somewhere in the, it's <u>somewhere in the front forwards</u> and…
> Paxman: is it <u>a set piece scrum or a ruck?</u>
> Johnson: it's <u>a set piece scrum and we're driving for the line and the ball's at our feet, and the enemy is wheeling, or trying desperately, pathetically, breaking the rules of the game, to…wheeling all over the place and we're heading, we're going for a push over try!</u>

The dialogue shows both interviewer and interviewee continuing the general conceptual metaphor POLITICS IS SPORT (studied for instance in Semino & Masci, 1996). Via references to specific aspects of the game of rugby, Johnson was able to level jibes at "the enemy" (the Labour party, presumably) and tactfully evade Paxman's real question of whether he (Johnson) was considering vying for the Prime Ministership.

Data from the research literature on L2 learners' tendency and capability to re-literalise idioms, and examples from popular culture are rare. One attempt to investigate this was made by Littlemore and Low (2006a, p. 131), who presented some of Prodromou's (2003) examples of manipulated idioms to advanced learners of English and asked them to come up with their adaptions to pre-specified contexts (Table 2.2):

Table 2.2 *Idiom Extension Task (Littlemore & Low, 2006a, p.131)*

Idiom	Adapt to the context of
1. To keep up with the Joneses	Tony Blair's positioning in relation to the USA.
2. To bring home the bacon	To refer to a person who earned a lot of money for their family.
3. Give him an inch and he'll take a yard	Someone who does this to excess.
4. He's a few sandwiches short of a picnic	A builder whose stupidity stops him from doing his job very well.
5. It's as easy as falling off a log	To refer to something that seems easy but isn't.
6. To stink to high heaven	To refer to something that smells extremely bad.

The learners produced a few acceptable adaptions: "bring home the dirty bacon" (idiom 2); "give him a hand and he will take your arm", "give him a drop of water and he will bring home the whole sea" (idiom 3); "a few pillars short of a house", "a few bricks short of a wall" (idiom 4) (p. 131). Additionally, there were several inappropriate responses: "to bring the boss a gift" (idiom 2); "it's not easy falling off a log" (idiom 5); "throw the drug away, it stinks to high heaven" (idiom 6) (p. 131). Importantly, the authors noted that on the whole, few learners were able to think of plausible adaptions, most found the activity very difficult, and the majority did not write anything.

This informal study highlights some of the practical issues of getting L2 learners to engage in metaphor-based language play. From it, Littlemore and Low concluded that while some (probably more advanced) L2 learners might benefit from and enjoy explicit discussions on the form of language play constructions and reasons why (or why not) re-literalised idioms are acceptable (Williams, 2001), such activities are likely to be very difficult for beginners, and some learners (whatever their general proficiency) may fail to see the point in relation to their own language learning goals.

2.2.3.7 …and issues facing L1 Chinese learners of English

L1 Chinese learners of English, such as the present study's participants, face several specific issues in their process of acquiring L2 English metaphoric competence.

Metaphor in phrasal verbs

Liao and Fukuya (2004) used a multiple-choice (gap-fill) test of tendency to select (literal and figurative) phrasal verbs over single word verbs and two distractors, a timed (L2 to L1) translation test, and a recall test to investigate English phrasal verb avoidance in L1 Chinese

intermediate and advanced learners of L2 English and an English native speaker comparison group. The findings showed that on a multiple-choice test, the intermediate learners selected (both literal and figurative) phrasal verbs much less frequently than both the advanced learners and the native speakers (suggesting they were avoiding them), and that on all three tasks the advanced learners used significantly more (literal and figurative) phrasal verbs than the intermediate learners. Avoidance was attributed to L2-L1 structural differences; Chinese has verb + particle structures, but particles are generally inseparable from the verb and few take on figurative meanings. Thus, the authors argued that the various semantic functions of English phrasal verb particles were likely to have been confusing to the intermediate L1 Chinese learners, leading them to avoid these forms.

Taking their data on L1 Chinese learners of English and combining it with patterns observed in L1 Dutch learners of English (Hulstijn & Marchena, 1989), Liao and Fukuya (2004) provided a model of the developmental shift from avoidance to nonavoidance of English phrasal verbs, arguing that regardless of L1 typological similarity with English, learners will go through the same process of avoidance to nonavoidance. Unfortunately, the fact that other similar studies have used L1 Hebrew (Dagut & Laufer, 1985) and L1 Spanish learners of English (Laufer & Eliasson, 1993) from the same proficiency level, means that the model cannot be retrospectively applied to this research. Limitations of Liao and Fukuya's (2004) study include the fact that metaphor was treated as an either/or phenomenon rather than as existing on a cline (Littlemore & Low, 2006a), and the fact that although definitions of *literal* and *figurative* from previous studies (e.g., Dagut & Laufer, 1985; Laufer & Eliasson, 1993) were given, the authors did not explicitly define these terms in their study.

Morphemes and phraseological accuracy

In their effort to attain L2 English phraseological proficiency (Philip, 2010), L1 Chinese learners of English are also likely to experience difficulties with English morphemes. In a meta-analysis of the influence of the L1 on the acquisition order of L2 English articles, plural s- and possessive s-, Luk and Shirai (2009) found that while L1 Spanish learners followed Krashen's (1977) Natural Order,[11] L1 Chinese, Japanese and Korean learners first acquired possessive -s, a form represented in these languages, and plural –s and articles later, forms not represented. Empirical association between absence of an L1 morpheme equivalent and inaccurate production in L2 English suggests that in written production, L1 speakers of Chinese may find metaphors and metonyms involving articles, plural -s, past tense -ed and third person -s particularly challenging to produce accurately. On the other hand, such learners (in theory) should have less trouble with

[11] Krashen (1977) proposed the following order of acquiring English morphemes: -ing / plural –s / copula be > auxiliary be / articles > irregular past tense > regular past tense / 3rd person –s / possessive –s.

the accuracy of those involving possessive -s and progressive -ing, since these are grammatically represented in Mandarin Chinese. These findings are relevant because several of the forms just mentioned are involved in a test developed for the present study to measure sensitivity to the acceptability of semantic and syntactic exploitations of Vehicle terms (Low, 1988; see also section 4.3.5.4 in the present study).

Chinese and English linguistic and conceptual metaphors: Similarities and differences

Many similarities and differences exist between Chinese and English linguistic and conceptual metaphors. Both languages appear to conceptualise LOVE as a JOURNEY, PLANT, FIRE, and UNITY, which may point to these being so-called primary (fundamental) metaphors stemming from embodied experience (Gibbs, 2011; Grady, 1997, 1999). In a study on anger metaphors in English-to-Chinese translations, Zhang (2013) provided several examples of corresponding English and Chinese linguistic metaphors for ANGER IS FIRE, ANGER IS COLOUR, ANGER IS KEEPING (FAILING TO KEEP) THE PRESSURE BACK, and ANGER IS A NATURAL PHENOMENON. For instance, both languages permit the conceptualisation of ANGER as RED and PURPLE, as evidenced in the following English-to-Chinese and Chinese-to-English literary translations: "Boxtel's face was red with anger / 波泰尔顿时气得脸色通红"; "老通宝气得脸都紫了 / the old man's face turned purple with rage" (p. 792).

Kövecses (2010) recorded that both Chinese and English share UP, LIGHT, and FLUID IN A CONTAINER as source domains for HAPPINESS, however, only Chinese has the metaphor HAPPINESS IS FLOWERS IN THE HEART and only (American) English has HAPPINESS IS BEING OFF THE GROUND. These differences are attributed to corresponding introverted and extroverted national characters (Kövecses, 2010; Yu, 1998). While one may object that such an interpretation betrays a stereotypical, overgeneralisation, Lv and Zhang (2012) concur that the Chinese-specific concepts LOVE IS SILK and LOVE IS (LIKE) THE MOON, and the English-specific concepts LOVE IS A COMMODITY and LOVE IS (LIKE) THE SUN reflect the broadly introverted, private-focused Chinese national character, and the broadly extroverted, public-focused English (or American) national character (cf. Su, 2002). Lv and Zhang also made the somewhat questionable assertion that cultural differences can account for the fact that "in Britain, people and dogs keep a close contact with each other, [and so] it is likely to find conceptual metaphors like DOGS ARE FRIENDS, such as in 'love me love my dog', 'you are a lucky dog'" (2012, p. 356). Despite the fact that much of this research has not employed robust (e.g., corpus-based) methods to arrive at conclusions, the general point is that cultural differences between Chinese and English may impact (positively or negatively) on the acquisition of L2 English linguistic and conceptual metaphors.

2.2.4 Research into L1 metaphoric competence

In contrast to L2 metaphoric competence research, L1 metaphoric competence appears to have largely focused on the processing of metaphor (rather than its production), child development, and developmental issues such as autism. In L1 (but not L2) metaphoric competence research,[12] a family of techniques known as Structural Equation Modelling (SEM), including Exploratory and Confirmatory Factor Analysis (EFA and CFA) and Principal Components Analysis (PCA), have been used to investigate whether scores for several metaphoric competence tests point to more fundamental underlying (sub)competences, and whether large numbers of variables can be parsed into smaller meaningful clusters. Because such techniques will be applied in the present study (the first time in L2 metaphoric competence research), two studies in which they have been used to investigate L1 metaphoric competence are synthesised for comparative purposes.

Replicating Pollio (1977), Pollio and Smith (1980) applied PCA to 70 participants' scores to a battery of metaphoric competence tests,[13] finding that 21 dependent variables were underpinned by four factors: Verbal Fluency (explaining around 50% of the total variance); Flexibility of Verbal Comparisons (explaining 21%); Logical Reasoning (explaining 11%); Innovative Figurative Use (explaining 10%). A fifth factor also emerged, but was reclassified as part of Innovative Figurative Use. Another PCA, this time involving 28 dependent variables, was conducted on the same 70 participants' data, and yielded five factors: Associative Fluency; Sensitivity to Poetic Diction; the Torrance Test Factor (Flexibility of Verbal Comparisons from earlier analyses); the Syllogisms Test Factor (Logical Reasoning from earlier analyses); Innovative Figurative Use. The authors concluded that analogy should be considered a special kind of metaphor, but not used as a general model for all figurative activity.

Limitations include the unaddressed issue of low sample-to-variable ratios of less than three participants per variable (A Field, 2013), the misuse of PCA, a technique for reducing or consolidate variables to identify "meaningful subgroups [and] explore the structure of verbal problem solving" (H. R. Pollio & Smith, 1980, p. 373) for which EFA would have been more appropriate[14] (Plonsky & Gonulal, 2015), and the lack of information about the participants' backgrounds, estimates of model adequacy and reliability of factors.

In a more recent study, Beaty and Silvia (2013) investigated the extent to which several

[12] Here, I refer to metaphoric competence as a collection of skills and abilities involved in real-word language use, however, it should be acknowledged that some studies on metaphor and brain activity have used Structural Equation Modelling techniques on L2 data.

[13] The MC Test Battery included tests measuring knowledge of analogies, adjective-noun associations (H. R. Pollio, Barlow, Fine, & Pollio, 1977; Stumberg, 1928), creative compositions, the Gardner Metaphor Preference Test (H. Gardner, Kirchner, Winner, & Perkins, 1975), logical syllogisms (Lefford, 1946), oxymorons, the Pollio Test of Metaphoric Comprehension (M. R. Pollio & Pollio, 1979a), similes, symbols, and the Torrance task ('unusual uses of an object' subtest) (1974).

[14] This consideration should be balanced by the fact that in 1980, the available software was significantly less powerful than today!

factors of the Cattell-Horn-Carroll model of intelligence contributed to the generation of L1 conventional and creative metaphors by 191 L1 English undergraduate participants. Conventional metaphor production was measured as "the ability to generate a vehicle term that aptly fits the constraints of an attributive category" (p. 259) via a timed, fill-in-the-blank task (taken from Chiappe & Chiappe, 2007) requiring participants to produce metaphors to describe entities such as boring jobs. Responses were scored for aptness by two raters using a six-point scale. Metaphor creativity was measured via a task requiring participants to describe two past experiences in a creative, clever, humorous, original, compelling or interesting way. Responses were scored on a five-point creativity scale, with Topic and Vehicle distinctiveness, novelty, and cleverness assessed. Fluid intelligence was measured via timed odd-one-out letter set, Cattell Culture Fair Intelligence, paper folding, and broad retrieval tasks. Crystallised intelligence was measured via vocabulary, general knowledge and personality tests.

The researchers took a robust approach to reliability, using generalisability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to compute interrater reliability on a Cronbach's alpha scale, and treating different raters' scores as different variables in subsequent correlations. Confirmatory Factor Analyses, conducted on 19 dependent variables before analysing the full structural model, suggested five underlying factors: Creative Metaphor Production; Conventional Metaphor Production; Fluid Intelligence; Broad Retrieval Ability; Crystallized Intelligence.

The findings of the main structural model extended those of past research (Silvia & Beaty, 2012), and showed that the quality of creative metaphors was best predicted by higher-order mechanisms associated with executive processes, involving the ability to simultaneously maintain an attributive (i.e., guiding semantic) category in mind, search semantic memory, and counter the interference of inappropriate lexical and semantic information. The ability to generate conventional metaphors, on the other hand, was best predicted by the crystallised knowledge tasks.

Despite their inherently different foci, two general lessons from these studies reviewed here can be drawn. First, because of the complexity of identifying latent (i.e., hidden) traits in observed scores, it is crucial to make informed methodological decisions. As highlighted, Pollio and Smith (1980) were much less robust in this respect than Beaty and Silvia (2013). Second, the factor structures of both studies revealed a general distinction between L1 metaphoric (sub)competences related to creative and conventional metaphor. The extent to which this distinction would also be seen in equivalent SEM of L2 data remains to be seen.

2.3 Vocabulary knowledge (L1 and L2)

2.3.1 What does it mean to know a word?

A fundamental part of second language learning is the acquisition of vocabulary knowledge. Definitions of vocabulary (or lexical) knowledge generally fall into either *trait* or *interactionalist* views (Laufer & Goldstein, 2004). The trait view conceptualises vocabulary knowledge as the sum of interrelated subknowledges (e.g., of written form, morphology, collocations), knowledge of the social constraints on word usage (Nation, 1990, 2001; J. C. Richards, 1976; Ringbom, 1987), and in terms of continuums such as receptive to productive knowledge (Palmberg, 1987). Most tests of vocabulary knowledge adhere to this view, thus targeting and measuring precise aspects of knowledge. By contrast, the interactionalist view rejects the premise of a decontextualised vocabulary knowledge, instead treating it as part of communicative competence (e.g., Bachman, 1990) and inseparable from other skills involved in interaction. While the interactionalist view taps into more real-world (i.e., non-laboratory) constructs, measurements using this approach are confounded by the inextricability of vocabulary knowledge from contextual factors such as strategy use, relationships between interlocutors, and other discourse dynamics.

2.3.2 What are 'receptive' and 'productive' vocabulary knowledge?

Vocabulary (and indeed language) knowledge is often discussed in terms of *receptive* and *productive* skills. Traditionally, reading and listening have been seen as receptive skills (involving input), whereas writing and speaking have been seen as productive skills (involving output). Alternatively, the notions of receptive and productive might be understood in terms of translation from the L2 to L1 (receptive) and L1 to L2 (productive), namely, from less-to-more or more-to-less familiar systems. A third way to understand receptive and productive knowledge has been in terms of *recognition* of features such as a word's form, meaning, sound or collocates (receptive) and being able to *recall* these when required (productive).

Although these conceptualisations seem straightforward, they are all problematic. One would struggle to write without reading, or speak without (in the broadest possible sense) listening. Similarly, the fact that any direction of translation (L2-L1, L1-L2, L3-L2, etc.) requires some degree of receptive and productive knowledge, and that L2-L1 translation has sometimes been reported as a productive (active) process complicates this account (Schmitt, 1999). Confusion around the recognition and recall distinction can be demonstrated by consideration that multiple-choice recognition tasks require test takers to recall the meanings of distractors, and recall tasks require recognition of a contextual meaning before a form can be recalled

(Melka, 1997; Webb, 2005).

So, is receptive and productive knowledge fundamentally distinct? Read (2000) suggests that if the answer to this question is *yes*, then research should seek to decipher at which point receptive knowledge becomes productive knowledge. Meara's (1997) theory on the network of word associations in the mental lexicon constitutes one response to this issue. This theory conceives of receptive knowledge as lexical items that have no connection to any others in the lexical network, and productive knowledge as lexical items that do. Lexical items that are part of receptive knowledge cannot be activated except via an external stimulus (e.g., encountering the word form when reading) whereas those that are part of productive knowledge can be activated by other connected items. In this view there is no natural time-dependent progression from a receptive to productive state, which may explain how some students seem to be able to learn words productively from seemingly little input and in a short space of time Schmitt (2000). In subsequent publications, Meara (2004, 2005, 2006) modelled the lexicon as a random autonomous Boolean network (rather like a series of interconnected on/off light switches) to investigate the processes behind language attrition and so-called *kick-starts* (i.e., bursts) in language learning. While these models have been useful for theorising, as computer simulations that do not necessarily reflect real world phenomena, they are fundamentally flawed (Schmitt, 2000). Criticism of this kind was made by Laufer (2005), who in response to Meara's use of Monte Carlo simulation to critique her Lexical Frequency Profile, contested that such models provide "a convenient escape from the real world, in which real people produce real language" (p. 587).

A further question for researchers concerns the extent to which the strength of relationship between L2 receptive and productive vocabulary knowledge changes as L2 proficiency increases. A study by Henriksen (2008) on L1 Danish adolescent learners of English sheds some light on this question. The author found no correlations between a receptive word connection task and a productive version of the WAT (see section 2.3.4.2) for any of the Grade 10-13 participants. However, large correlations (significant at the .01 level) were found between the productive WAT and a receptive vocabulary size test (the VLT from Schmitt et al., 2001), decreasing from Grade 7 participants (r = .85) through to those in Grade 10 (r = .69) and Grade 13 (r = .55). These findings appear to suggest that the strength of the relationship between receptive and productive vocabulary knowledge decreased as L2 proficiency (in this case, indexed by grade/age) increased.

According to some, this finding is intuitive to what would be expected. Schmitt (2010) suggests that L2 learners would first establish meaning recall, then start to develop other aspects such as grammatical and morphological knowledge, which would facilitate receptive recognition during reading and listening, in the end acquiring form recall. Crucially, at this latter stage,

learners would require "more time to fill in the contextualized elements of word knowledge (e.g., collocation, register) to a point where the lexical item could be confidently used in an appropriate manner in a variety of spoken and written contexts" (p. 87). The implication, therefore, is that the strength of the relationship between receptive and productive vocabulary knowledge would weaken at higher L2 proficiency levels, because form recall (productive knowledge) necessarily progresses more slowly than the rate at which learners recognise at least one (minimum) meaning aspect of new forms. Laufer and Goldstein (2004) demonstrated that L1 Hebrew, Arabic and Russian learners of English[15] showed the following order at all L2 proficiency levels:

1) Form recall[16] (the most difficult skill): Supply the L2 target form from the target form in the L1 and an initial letter prompt in the L2;
2) Meaning recall: Supply the L1 form from the target L2 form;
3) Form recognition: Choose the correct L2 target form (from four options) given the L2 target meaning;
4) Meaning recognition (the easiest skill): Choose the correct L2 target meaning (from four options) given the target L2 form.

These findings imply that the ability to *recognise the meaning* of, say, 1,000 more words would mean the ability to *recognise the forms* of somewhat fewer, to *recall the meanings* of yet fewer, and *recall the forms* of even fewer than that. In a regression analysis, the authors also found that knowing the form-meaning link of words accounted for 42.6% of the total variance in participants' class grade scores, with meaning recall the best predictor. The fact that these scores were assessed via components of reading, listening, speaking, writing, grammatical accuracy, sociolinguistic appropriateness, and language fluency suggests that vocabulary knowledge is likely to contribute a very large amount to overall success in a second language (Schmitt, 2010).

Other data on L2 receptive and productive vocabulary knowledge in relation to L2 proficiency come from the vocabulary strand of DIALANG (Alderson, 1995), a large project for the development of diagnostic language tests in 14 European languages. This research showed that productive vocabulary measures such as gap-fill and word formation (e.g., when initial letters and context are given) had consistently higher correlations with the proficiency components than receptive vocabulary tests such as meaning recognition and collocation recognition did. While all vocabulary measures correlated strongly with all second language proficiency components (r = .61 to .79), correlations with writing were strongest. In these tests,

[15] The L1 Russian participants were also first language speakers of Hebrew, and so their L1 tasks were in Hebrew.

[16] Schmitt's (2010) easier to understand reformulation of Laufer and Goldstein's (2004) original terms (in parenthesis) are used: 1) form recall (active recall); 2) meaning recall (passive recall), 3) form recognition (active recognition), 4) meaning recognition (passive recognition).

vocabulary knowledge accounted for 37-62% of the variance in the various language proficiency scores. Schmitt (2010) notes that, given all the factors that might contribute to L2 proficiency (e.g., motivation, background knowledge, familiarity with test task), it is remarkable that one factor could account for such a large amount of variance in proficiency scores. These findings appear to show that "language ability is to quite a large extent a function of [receptive and productive] vocabulary size" (Alderson, 1995, p. 88).

Webb (2005) also investigated differences between L2 receptive and productive vocabulary knowledge. In this study, the author engaged 66 and 49 L1 Japanese learners of English in receptive (sentence gloss reading) and productive (sentence writing) treatment tasks to observe their effect on receptive and productive vocabulary knowledge, measured via tests of knowledge of orthography, meaning and form, grammatical functions, syntax and association. The first experiment showed that the receptive treatment task was superior to the productive treatment task for learning both of receptive and productive vocabulary when all participants were permitted the same amount of time to complete tasks (12 minutes). In a second experiment, when the duration of tasks was not restricted, the productive (sentence writing) task led to greater receptive and productive vocabulary gains. Webb's experiments showed that the effectiveness of receptive and productive learning depends crucially on time allocated. To the extent that conditions in the second experiment are more reflective of those in an EFL classroom, the results argue for the use of productive (over receptive) learning tasks.

Concerning L2 metaphoric competence, research to date offers only snapshot correlations between receptive and productive measures at intermediate levels (Azuma, 2005; Littlemore, 2001); these cannot provide information about any change in correlation strength depending up L2 proficiency level. To the extent that L2 receptive and productive metaphoric competence behave like receptive and productive vocabulary knowledge, the strength of their relationship would be expected to decrease as L2 proficiency increases. This, however, is an entirely open question.

2.3.3 What affects L2 vocabulary learning?

Several factors are thought to affect the learning of vocabulary, including the frequency of a word in language, the number of times and ways in which it is encountered by learners, patterns of phonemic, phonotactic and derivational regularity, deceptive morphological transparency, word length, part of speech, concreteness and abstractness (Laufer, 1997). On the issue of frequency, Laufer and Goldstein's comment that "the distractors are taken from the frequency level of the target word, which makes them as difficult for the learner as the target word" (2004, pp. 406-407) seems to suggest that lower frequency implies difficulty. But is this true?

While many high frequency words may be more easily learned, Gass and Mackey (2002)

note that some acquisition appears to proceed regardless of frequency in the input, for instance, when fixed L2 developmental sequences such as those relating to morphemes are involved, or when negative evidence (i.e., examples of what is unacceptable in a language) is required. Concerning the latter issue, Trahey and White (1993) observed that L1 French child learners of L2 English who were exposed to lots of examples of Subject-Adverb-Verb order (e.g., 'she always runs fast') were able to notice that this combination is possible in English, but importantly, did not learn that Subject-Verb-Adverb-Object (e.g., 'she eats always chocolate') is not. In order for learners to learn that such forms are non-targetlike, negative evidence is required. Similarly, in a study involving UK- and Poland-based L1 Polish learners of English, Foster et al. (2014) investigated the influence of a variety of independent variables on nativelike selection (NLS), namely the ability to recognise word combinations as native or non-nativelike. Using a group of English native speakers as a baseline, the authors found that NLS equivalent to English native speakers was only attainable by UK-based non-native speakers who had started learning English before the age of 12 years old. For late starters, a good phonological short-term memory accompanied by immersion exposure predicted NLS in late starters and brought some gains (though not to nativelike levels), whereas positive feelings toward English and motivation to interact had no relationship with NLS. The study essentially showed that a good phonological short-term memory and immersion are necessary conditions for acquiring a nativelike ability to recognise word combinations as either native- or non-nativelike.

2.3.4 Vocabulary size and depth (L1 and L2)

A crucial distinction in vocabulary knowledge research has been made between the number of forms a language user knows (vocabulary size) and the quality of knowledge that the language user has of these forms (vocabulary depth). Schmitt (2010) and others (e.g., Goulden, Nation, & Read, 1990) have concluded that most educated native speakers are likely to *know* (in the broadest possible sense) 16,000-20,000 word families. For non-native speakers, 2,000-3,000 word families are required to be able to hold conversations in English if 95% coverage is needed, and between 6,000 and 7,000 if 98% coverage is needed (Schmitt, 2010). Webb and Rodgers (2009a, 2009b) found that L2 learners of English need 6,000-7,000 words to watch movies, and 5,000-9,000 to watch television. In the written mode, 8,000-9,000 word families are needed for reading texts such as novels and newspapers, assuming 98% coverage (Nation, 2006). As a rule of thumb, Nation (2001) notes that the first 1,000 most frequent words will make up around 70-75% of a typical text, with the next most frequent 1,000 accounting for an extra 5-8%. For academic texts, this figure is likely to be much higher at perhaps 20,000 word families needing to be recognised in order to read an academic text comfortably (Nation & Webb, 2011).

However, methodological flaws in much vocabulary size research, genre or profession-

specific terminology, and the fact that a higher level of education may not always go hand in hand with a higher vocabulary size problematise these generalisations. A crossword enthusiast who left school at 16, for instance, may know more words than an established academic.

2.3.4.1 Tests of vocabulary size

Yes/No (checklist) tests

The Yes/No (checklist) vocabulary test format is simplest for measuring how many words a language user has receptive knowledge of. Tests using this format such as the Eurocentres Vocabulary Test (Meara & Jones, 1990) and V_YesNo v1.0 (Meara & Miralpeix, 2015, hereafter VYesNo) tend to consist of real words selected randomly from various frequency ranges, and pseudo-words created via a random assortment of syllables from the words in these frequency ranges. Pseudo-words are then checked by native speakers to ensure they align with the phonological rules of English (Huibregtse, Admiraal, & Meara, 2002). All words are presented to test takers, who are required to indicate (via the click of a button) whether they recognise the word as a real English word.

Advantages of the Yes/No format stem from its efficiency, allowing for the administration of many items to many participants, and the relative ease with which such tests can be developed and scored (Pellicer-Sánchez & Schmitt, 2012). Yes/No tests enjoy high validity, and are strongly correlated with other tests of L2 receptive vocabulary size but less so with L2 productive vocabulary knowledge measures (Anderson & Freebody, 1983; Carey & Harrington, 2009; Pellicer-Sánchez & Schmitt, 2012). Meara and Buxton (1987) found that the one particular vocabulary size test using a Yes/No format (the Eurocentres Vocabulary Test) was better at discriminating between test takers than similar tests using the multiple-choice format (see below).

Vocabulary Levels Test (VLT)

The Vocabulary Levels Test (VLT) (devised by Nation, 1983) and its derivatives use multiple-choice format, and are probably the most widely administered tests of vocabulary size to date. The original VLT was a diagnostic instrument divided into five parts corresponding to frequency bands of the most common 2,000, 3,000, and 5,000 English words, and those beyond the 5,000 (University Word Level) and 10,000 most common. Each level lists 36 words and 18 definitions in groups of six and three respectively, and requires test takers to match the three target words (on the left) with correct definitions (on the right). For example, test takers must match ‘apply’, ‘elect’, ‘jump’, ‘manufacture’, ‘melt’, ‘threaten’ with ‘choose by voting’, ‘become like water’, and ‘make’ (Read, 2000, p. 119).

All words in each group belong to the same grammatical class, so as not to provide

grammatical clues. Words of similar meaning are not grouped together and thus the test is designed to be a measure of broad knowledge rather than of subtleties between semantically related words (Read, 2000). A major limitation of the VLT format is that it is possible for a test taker who knows the target word to get it wrong because they do not know words contained within the context or definitions, or to know an aspect of the target word's meaning not targeted by the definition (Meara & Buxton, 1987). Another problem is that as vocabulary size increases, so does the number of items that need to be tested if the proportion of words known is to be kept the same. For instance, if a learner who knows 1,000 words is given a test of 25 items, one in every 40 words that they know are tested. For a learner who knows 10,000 words, the same test targets only one in every 400 words that they know.

A vocabulary size test of controlled productive ability

Laufer and Nation (1999) developed a vocabulary size test of productive ability in a *controlled* (i.e., elicited) context. This test provided a useful compliment to their previously developed Lexical Frequency Profile (Laufer & Nation, 1995), a measure of the lexical richness of vocabulary in free, naturalistic production.[17] Resembling the C-test format (Klein-Braley, 1985; Klein-Braley & Raatz, 1984), each item in the authors' controlled productive ability test contains a meaningful sentence with an incomplete/mutilated word used to ensure test takers do not fill in the gap with a semantically appropriate (non-target alternative). For instance, 'the garden was full of fra____ flowers' aims to elicit the production of 'fragrant' via the provision of the initial three letters (Laufer & Nation, 1999, p. 33). From the data collected, the authors concluded that the test is a reliable, valid and practical measure of controlled productive vocabulary and suggested using it alongside the receptive VLT and Lexical Frequency Profile to investigate further questions about the development of L2 vocabulary knowledge and relatedness of its components.

2.3.4.2 Tests of vocabulary depth

Vocabulary knowledge Scale (VKS)

The Vocabulary Knowledge Scale (VKS) (Wesche & Paribakht, 1996) is one measure of how well learners know the words that they know. The VKS requires test takers to supply ratings of how well they know the meaning of target words, and for any words claimed to be known well, provide a synonym, translation or example of the word in a sentence. While its format allows for robust measurement, problems concern the high burden placed on test takers, and subsequent limitations in the number of items that can be administered. Furthermore, since the

[17] Since elicited metaphor, rather than metaphor in free, naturalistic production is the subject of this thesis, research into type-token ratios and other aspects of lexical diversity and richness is not reviewed.

test relies on test takers being able to clearly articulate word meanings, it risks eliciting ambiguous (i.e., difficult to score) productions.

Word Associates Test (WAT)

In the realisation that the Yes/No format (see above) lacked a measure of how well learners know the words they purport to know, Read (1993) developed the Word Associates Test (WAT) to measure the quality (or depth) of receptive lexical knowledge construed as the ability to identify semantic and collocation associations (i.e., a receptive test of vocabulary depth). The WAT presents learners with stimulus adjectives (50 in earlier versions, 40 in later versions) and eight possible associates (four adjectives on the left and four nouns on the right). Four of the eight words relate to the stimulus in one of three ways: (1) as paradigmatic adjective associates (words on the left) synonymous or at least similar in meaning to the stimulus adjective, perhaps with one being more general than the other; (2) as syntagmatic noun associates (words on the right), namely frequent collocates with the stimulus adjective; (3) as analytic adjective associates (words on the left) representing one aspect or component of the meaning of the stimulus word, likely to form part of its dictionary definition.

The WAT thus conceptualises vocabulary depth in terms of the degree to which any item is linked to other words in the mental lexicon, or lexical organisation (section 2.3.2). To date, the lexical organisation conceptualisation has been the most commonly used format for measuring vocabulary depth (Schmitt, 2014).

Read (1993) developed the WAT with two key assumptions: that native speakers "have" (p. 358) and presumably *produce* stable patterns of word association reflecting their rich lexical and semantic networks; and that second language learners "produce" (p. 358) fewer stable associations, with those they do produce tending to be based on phonological rather than semantic links with stimulus words. As proficiency increases, the author argued, non-native speaker instability tends towards stability. Fitzpatrick (2007) has extended the investigation into receptive vocabulary depth (measured by the WAT) to productive vocabulary depth, challenging the pre-existing notion of native speaker stability. In an experiment involving lower frequency words and non-concrete noun stimuli (both known to produce more predictable responses) the author found that 30 adult English NSs were not homogenous or predictable in their response behaviour as a group, with large discrepancies between participants. One participant, for instance produced as many as 57 consecutive collocations whereas another produced only five. It was also found that many participants had discernible, predictable response types, producing for instance mainly meaning based associations. These points serve to highlight that native speaker word association stability varies considerably when the receptive-productive, high-low word frequency and concrete-abstract continuums are altered. The real question for non-native

speakers, she argues, is not whether they move towards a native, norm-like profile as L2 proficiency increases, but whether their own individual profile becomes more established.

1K-Vocabulary Depth Test

Richard's (2011) 1K-Vocabulary Depth Test (1K-VDT) presents another measure and format for testing vocabulary depth. In this test, test takers are required to: (1) decipher the target word from gaps in six example sentences corresponding to its different dictionary definitions;[18] and (2) once known, supply the word in each sentence, correctly attending to aspects such as morphosyntactic form required. For example (p. 118):

(Answer = *arm*)

1. She held the young boy in her ___[arms]___
2. Matsuzaka [Japanese baseball pitcher] has a good ___[arm]___
3. As they walked, he offered her his ___[arm]___
4. The political ___[arm]___ of the group met with the media.
5. Both sides agreed to ___[disarm]___
6. Mom ___[armed]___ us with supplies to get the house ready.

Thus, the 1k-VDT measures productive knowledge of grammatical structures, affixes, collocations, and phrase-based usage. Richards validated this test by establishing its correlation with Nation and Beglar's (2007) Vocabulary Size Test (VST), a reading test requiring test takers to match forms with meanings via four-option multiple-choice. Advantages of the 1k-VDT relate to the scope of knowledge engaged (e.g., collocational, grammatical, semantic), and disadvantages concern the reading burden (and thus coverage restriction) and the fact that the test does not capture information about which clues test takers actually found most useful.

Webb's (2005) 10-pronged approach to vocabulary knowledge (involving depth)

Webb's (2005) 10-pronged approach stands apart as a highly innovative approach to vocabulary knowledge measurement. In this study, knowledge of each target word was measured in 10 different ways, through tests of receptive and productive knowledge of orthography, meaning and form, grammatical functions, syntax and association. A multiple-choice format was used for receptive tests of orthography, syntax, association, and grammar whereas the receptive test of meaning employed an L2 to L1 translation format. For all of the productive tests, participants were presented with decontextualised cues and required to produce a response to demonstrate the aspect of knowledge being measured. In order to ensure that learners did not have any prior knowledge of the target items, nonsense words matched with the meanings of low frequency

[18] Richard's (2011) used Collins Cobuild Dictionary definitions with low frequency words replaced by synonyms within the 1000 most common English word range.

English words were developed.

2.3.4.3 What is the difference between size and depth of vocabulary knowledge?

Enquiry into this question dates back to at least Anderson and Freebody (1981). Its importance can be seen by considering the relevance of vocabulary size and depth to EFL teachers seeking to understand why some students seem to know very few words but know them well, while others recognise a large number of words but do not know much about them (Schmitt, 2014). For some, the high correlation between measures of vocabulary size and depth indicate that they are fundamentally the same construct. Vermeer contested that "a deeper knowledge of words is the consequence of knowing more words, or that, conversely, the more words someone knows, the finer the networks and the deeper the word knowledge" concluding "there seems to be no conceptual distinction between breadth [size] and depth" (2001, p. 222).

Others object. Gyllstad (2013), drawing on Meara and Wolter's (2004) argument, suggests that vocabulary size is a measure of a learner's entire vocabulary and as such is not a characteristic of individual words. Vocabulary depth, he suggests, *is* typically a characteristic of individual words, making extrapolation to other words impossible "or at the very least, difficult" (p. 19). Put simply, Gyllstad's point is that vocabulary size and depth are conceptually distinct because it is impossible to predict the degree to which a learner knows a word from measurements of how well they know other words. Schmitt (2014) seems to agree, concluding that the extent to which vocabulary size and depth are separate entities depends on how they are conceptualised and measured (e.g., as knowledge of multiple aspects of words, polysemous meaning senses, derivative forms, collocation, lexical fluency or lexical organisation). From a lifetime and career spent researching vocabulary knowledge, Schmitt (2014) suggests that for him, the concept of lexical organisation appears to provide the most promising approach for future research into vocabulary size and depth.

Empirical research on vocabulary size and depth, and their relation to L2 proficiency offers a number of findings. First, both vocabulary size and depth are basically indistinguishable as predictors of L2 reading comprehension, though depth may have a slight upper hand. In a study on 74 L1 Chinese and L1 Korean learners of English, Qian (1999) observed that L2 vocabulary size was the best predictor of L2 reading comprehension, with L2 vocabulary depth making a significantly additional (but smaller) unique contribution. In a subsequent to study on L2 learners of English from mixed L1 backgrounds[19], Qian (2002) also found depth of vocabulary knowledge (DVK) (the author's adaption of Read's 1993 WAT) to be a slightly stronger predictor

[19] Participants were L1 Korean, Japanese, Spanish, Chinese, Tajik, Arabic, Portuguese, Russian, Italian, and from 10 other languages.

of TOEFL Reading for Basic Comprehension (RBC) scores (R^2 = .59, p < .01) than both the VLT (Nation, 1983), a vocabulary size measure, and TOEFL Vocabulary Item Measure (R^2 for both = .54, p < .01). Using a step-by-step hierarchical regression, Qian found that the VLT (vocabulary size) provided a significant additional 8% of the criterion variance over and above the DVK (vocabulary depth), whereas The DVK (vocabulary depth) provided a significant extra 13% over and above the VLT (vocabulary size). Similar significant sizes of variance were found in hierarchical regression analyses involving other predictors, suggesting that using any combination of the VLT, DVK and TOEFL-VIM led to better predictions of L2 reading comprehension than any one alone. These studies show that despite some nuances, vocabulary size and depth can be considered equivalent at predicting L2 reading comprehension (Schmitt, 2014). However, the extent to which vocabulary size and depth predict L2 receptive and productive metaphoric competence is unknown.

A second finding is that SEM suggests that vocabulary size is (slightly) more central to the construct of vocabulary knowledge than vocabulary depth (construed as lexical organisation)**.** In an analysis using SEM to draw out latent underlying variables from observed measures, Zhang (2012) found a latent variable vocabulary knowledge, and that the VLT loaded strongly onto this variable (β = .86), whereas a Word Associates measure had a weaker loading (β = .60). This results are corroborated by Tseng and Schmitt (2008), who also found that a vocabulary size measure loaded more strongly (β = .71) on a latent vocabulary knowledge variable than a vocabulary depth measure did (β = .67).

Third, there are mixed findings on whether vocabulary size and depth (construed as lexical organisation) converge or diverge in strength as L2 proficiency increases. Studies on Grade 7 to Grade 13 L1 Danish learners of English (Henriksen, 2008) and L1 Japanese learners of English (Noro, 2002) have found that vocabulary size and depth (measured using VLT and the Word Associates Test), appear to be more highly correlated for learners with smaller vocabulary sizes and for high frequency words, growing further apart as proficiency increases, with depth seemingly lagging behind size at higher levels. These studies seem to imply that the form-meaning link is easier than the type of lexical organisation measured by the Word Associates Test. Other studies show the opposite pattern. In research using translation as the vocabulary size measure and the Word Associates format as the vocabulary depth measure, Nurweni and Read (1999) found higher correlations across three increasingly more proficient groups of L1 Indonesian learners of English. In a longitudinal study on L1 Japanese learners of English, Schmitt and Meara (1997) found increased correlations of vocabulary size and word association recall (r = .49 to .62) and recognition (r = .39 to .61) between tests administered at the start and the end of the school year (with average vocabulary sizes increasing from 3,900 to 4,230 word families).

Fourth and finally, research shows that vocabulary depth (construed as lexical

organisation) has the strongest correlation with form-recall, the most difficult aspect of vocabulary knowledge to acquire. In a study on L1 Dutch learners of French, Greidanus, Bogaards, van der Linden, Nienhuis, and de Wolf (2004) found that form-recall scores correlated with the Word Associates Test more highly (r = .81) than form-recognition did (r = .70). Such results demonstrate that lexical organisation seems to be related to the highest (i.e., most difficult to acquire) level of form-meaning knowledge, form-recall, as identified by Laufer and Goldstein (2004).

2.4 Language proficiency (L1 and L2)

2.4.1 L1 proficiency

Given the multitude of ways in which human beings are exposed to language(s) in their lifetimes, and the fact that multilingual speakers probably outnumber monolingual speakers, defining the concept of first (native, L1) language proficiency is highly problematic. Hulstijn (2011) proposed that native speakers of a language demonstrate two kinds of language ability: basic language cognition (BLC) and higher language cognition (HLC), which he argues can account for the fact that native speakers all share some aspects of language knowledge, but differ greatly with regard to others.

In this view, BLC denotes what all native speakers have in common, and concerns implicit knowledge domains such as phonetics, prosody, phonology, morphology and syntax, the explicit knowledge domain of lexis, and (though it may differ between speakers) automatic processing of these domains. BLC is restricted to frequent lexical items and grammatical structures that any normally developed[20] adult language user would be able to understand or use in the spoken mode. HLC, on the other hand, is the domain where differences between native speakers can be observed. These differences are caused mainly by varying "intellectual profiles" (Hulstijn, 2012, p. 428), an arguably contentious term[21] referring to intellectual skills, level of education, occupation and leisure time activities. HLC is identical to BLC except it concerns low frequency lexical items or grammatical structures, and pertains to both written and spoken modes. The author is careful to point out that morphological and syntactic structures, words and expressions cannot be categorised as either BLC or HLC on the basis of a strict frequency boundary, but that the issue, rather, is one of prototypicality.

Hulstijn's distinction between BLC and HLC leads to three hypotheses, the second of which states:

[20] Although not the concern of the present study, it should be acknowledged that BLC may manifest differently in people affected by serious language-related mental disorders (Hulstijn, 2012).

[21] Despite its overly simplistic and potential derogatory connotations, for convenience Hulstijn's term is used in this thesis.

> H2: Individual differences among adult L1-ers will be relatively large in tasks involving HLC discourse, in all four modes of language use (reading, writing, listening, and speaking) but almost all adult L1-ers will perform at ceiling in BLC tasks, that is, conceptually simple oral tasks (listening and speaking) involving highly frequent linguistic units. (2011, pp. 231-232)

This prediction has intriguing implications for L1 metaphoric competence, because it suggests that if a linguistic metaphor is comprehensible or producible in the spoken mode by all native speakers (higher and lower intellectual profiles) it is more prototypically part of BLC, whereas linguistic metaphors in the written mode which native speakers differ in their knowledge of, are more characteristic of HLC. Construed in this way, BLC and HLC give potential theoretical grounding to Low's (1988) observations (and assertions) about native speaker variation with regard to the acceptability of different metaphorical words and structures.

2.4.2 L2 proficiency

2.4.2.1 Models and frameworks

In a period spanning over 50 years, theories on the nature of L2 language learning have progressed from early two-dimensional grid models of linguistic knowledge and the four skills (e.g., Lado, 1961) to models of communicative competence involving: the possibility, feasibility, appropriateness, and actual usage of forms, phrases, structures (Hymes, 1972); grammatical, sociolinguistic and strategic competence (Canale & Swain, 1980); language knowledge (itself comprised of organisational and pragmatic competences) and strategic competence involving metacognitive components and strategies (Bachman, 1990; Bachman & Palmer, 1996), which some (e.g., Hulstijn, 2011) suggest are peripheral to language competence. Despite increasing recognition of the importance of communicative functions of language, empirical research (e.g., Bachman & Palmer, 1982; Harley, Cummins, Swain, & Allen, 1990; Sasaki, 1993) has had considerable difficulty confirming the hypothesised structures of language proficiency.

Another central question on L2 proficiency concerns whether or not post-puberty (i.e., late) L2 learners can acquire the target language to a nativelike level. In its strong form, what came to be known as the critical period hypothesis (Lenneberg, 1967; Penfield & Roberts 1959) predicts that late L2 learners can never achieve nativelike mastery. Weaker versions suggest such mastery is unlikely but nevertheless possible, but perhaps involving fundamentally different mechanisms for early and late L2 learners (Andringa, 2014). While late L2 learners can certainly acquire HLC, the extent to which they can master BLC remains an open question (Hulstijn, 2011).

2.4.2.2 Measurement scales

Stakeholders such as universities, businesses, border and immigration agencies and (not least) language learners themselves frequently require verification that a language has been learned sufficiently for a specific purpose. In the UK, where this research takes place, students whose first language is not English wishing to study in higher education usually take the International English Language Testing System (IELTS) exams. IELTS uses its own scoring scale, which can (in theory) be converted to CEFR levels for comparison with other measures such as the OOPT. Unfortunately, despite their usefulness for helping teachers and testers make practical decisions, all scales and tests of L2 proficiency can be criticised on several grounds.

The main problem seems to be the fact that L2 proficiency scales and tests inherently conflate language development with language proficiency. Thus, attainment of the CEFR levels B2, C1 and C2, for instance, requires not just higher language skills, but higher so-called 'intellectual skills' (Hulstijn, 2011). Specific tests such as IELTS tend to come under scrutiny for encouraging a view of language as an abstract, objective and context-independent entity, rather than inextricable from specific genres and subjects (Pilcher & Richards, 2016, 2017; K. Richards & Pilcher, 2016). Critics argue that gatekeepers deciding which non L1 English international students can access particular higher education institutions should give more weight to subject specialists, and less to exams such as IELTS (Pilcher & Richards, 2017). In the remainder of this section, two commonly used, standardised tests of L2 proficiency are examined more closely.

2.4.2.3 The Oxford Online Placement Test (OOPT)

The OOPT is, strictly speaking, a *placement* test used to obtain a quick and reliable measure of a student's general language ability for placing them at a particular level, although its robustness also makes it a useful measure of L2 proficiency. The test was developed with a mandate to measure more than knowledge of grammatical form, be short, straightforward and reliable, provide detailed performance feedback, and be capable of being customised (Purpura, 2009).

The test contains two parts: (1) a Use of English section to measure knowledge of grammatical forms, semantic meaning, grammatical form *and* meaning, and knowledge of pragmatic meanings encoded in social interactions; (2) a Listening section to measure ability to understand both the literal meanings encoded within the listening text, and implied or intended meanings encoded either within the text or beyond the parameters of the actual text.

The test uses CEFR descriptors as a point of reference for what students might be able to do with the language (not what they know) at different proficiency levels, thus measuring language knowledge and listening ability at one of six CEFR levels (A1 to C2). In the Use of English section, multiple-choice and fill-the-gap item foci include numerous components of language knowledge (e.g., noun phrases, tense and aspect, modals and phrasal models, phrasal verbs,

prepositions, conditionals, adverbials, reported speech). In the Listening section, test takers answer around 15 multiple-choice questions in response to short and longer dialogues or monologues tapping into these various language components. The administrator can choose whether the dialects of speakers are British, American, or both. The OOPT is computer adaptive and selects questions from a large bank of standardised items, which have been extensively piloted for reliability. If a previous question is answered correctly, a more difficult question follows and vice-versa. Students typically finish the test in 30-40 minutes, but can be permitted up to 90 minutes. Scores range from 0 to 120, with 20 points corresponding to each CEFR level.

2.4.2.4 International English Language Testing System (IELTS)

IELTS is a test of language proficiency for people seeking to study or work where English is used as a language of communication. IELTS tests Listening, Reading, Writing and Speaking abilities and uses a nine-band scale to identify levels of proficiency, from non-user (band score 1) through to expert (band score 9). The test is available in Academic or General Training versions which have identical Listening and Speaking but different Reading and Writing sections, and purports to avoid cultural bias, by including all "standard varieties of native speaker English, including North American, British, Australian and New Zealand English" (2017).

The Listening section of IELTS (30 minutes) contains four sections with 10 questions each, including dialogues and monologues. Task types include multiple-choice, matching, plan/map/diagram labelling, form/note/table/flow-chart summary completion, and sentence completion. The (Academic) Reading section (60 minutes) contains three passages and a total of 40 questions. Task types are similar to those in the Listening section. The (Academic) Writing section (60 minutes) contains two tasks/questions requiring a 150-word description of visual information (e.g., a graph)/table/chart/diagram), and a 250-word response to a point of view, argument or problem. Finally, the Speaking section takes the form of a recorded oral interview (11-14 minutes) between the test takers' and an examiner involving scripted questions about familiar topics such as home, work and family (4-5 minutes), a cue card, short talk and follow up questions and answers (3-4 minutes) and a freer, more abstract discussion of issues related to the talk (4-5 minutes). Certified IELTS examiners assess both the Writing and Speaking responses according to aspects such as coherence, lexical resource, grammatical range and accuracy, and fluency and pronunciation (speaking only).

2.4.2.5 Metaphor in the OOPT and IELTS

The first research paper on the OOPT website has seven mentions of the word 'metaphor' or 'metaphoric', four more than in the CEFR in fact (Nacey, 2013). One mention is as part of the sociocultural competence component of Purpura's (2004) earlier framework (Purpura, 2009, p.

6). The remaining six are as part of the descriptors for CEFR C2 mastery, “uses accurately with precision a wide range of vocabulary for unfamiliar and abstract topics. Can use metaphoric language idioms and colloquialisms and can convey finer shades of meaning.” (p. 13), and C1 mastery, “uses accurately and appropriately a wide range of vocabulary for unfamiliar topics. Can also use some metaphoric language and idioms.” (p. 13). Apparently then, the author does not consider metaphor to be an important part of CEFR A1 to B2 mastery measured by the OOPT. By contrast, a search of the IELTS website does not retrieve any result for ‘metaphor’ or ‘figurative language’, suggesting that not even IELTS scoring descriptors for higher levels explicitly acknowledge its role in L2 proficiency. On the other hand, linguistic metaphor might be assumed to form a part of the ‘lexical resource’ scoring component (Writing and Speaking). Despite some research into the extent to which IELTS taps into *pragmatic* knowledge and its (sub)competences (Allami & Aghajari, 2014), there does not appear to have been any investigation into L2 metaphoric competence and IELTS.

2.4.3 Formulaic sequences (L1 and L2)

Language that is metaphorical is largely and often undoubtedly ‘formulaic’ (contains some degree of fixed patterning, roughly speaking). Unfortunately, precise and consistent definition and identification of formulaic sequences (both metaphorical and non-metaphorical) are significant challenges, problematised by numerous overlapping and sometimes contradictory terminology (Myles & Cordier, 2017; Wray, 2002). Two general approaches to the identification formulaic sequences exist: speaker-external and speaker-internal.

Speaker-external approaches may include the identification of sequences via native speaker intuition and shared knowledge, their frequency (e.g., in corpora), grammatical structure, and (in spoken production), features such as phonological and fluency-based markers, stress and articulation. Several advantages and disadvantages with these approaches can be observed. Identifying formulaic sequences based on intuition roots the process in the perceptions of real language users, but coders suffer from lapses in concentration and inter- and intrapersonal inconsistency. With a frequency-based approach involving language corpora, a computer can perform consistent and high-speed identification without getting tired, but problems concern certain genre-specific forms being under or overrepresented, sequences being undetectable due to low frequency (e.g., ‘long live the king!’), the need for researchers to make numerous post-hoc decisions about irrelevant or uninteresting search results, and the fact that while a corpus may be broadly representative of language within the domain of its parameters, it cannot truly mirror the experience of an individual person or reflect language in certain domains (Schmitt, 2010).

Unsurprisingly, a wealth of research has found knowledge of formulaic sequences

(identified using speaker external methods) to be a strong predictor of general L2 language proficiency (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Dai & Ding, 2010; Hsu & Chiu, 2008; Keshavarz & Salimi, 2007; Stengers, Boers, Housen, & Eyckmans, 2011). Nevertheless, idioms, idiomatic expressions, collocations and lexical bundles are difficult for non-native speakers to master, even at the advanced level (Myles & Cordier, 2017). Wray (2012) agrees that instructed L2 learners, on the whole, have an "impoverished stock" (p. 236), which raises the question as to why L2 learners do not seem to capitalise on formulaic sequences in their L2 input. Major findings include the underuse, overuse and misuse of collocations, that non-native speakers tend to use more lexical bundles as discourse markers in writing than NSs, and that the influence of the first language seems to account for a large number of non-targetlike collocations (Paquot & Granger, 2012). One concerning finding from Boers, Demecheleer, Coxhead and Webb (2014), is that while South East Asian learners were better at correctly matching verb-noun collocation with their appropriate nouns if they had engaged in exercises in which these were presented and manipulated as intact wholes, gains were offset by the fact that learners had acquired some of the distractors in the exercises.

Speaker internal approaches define a formulaic sequence as "a multiword semantic/functional unit that presents a processing advantage for a given speaker, either because it is stored whole in their lexicon or because it is highly automatised" (Myles & Cordier, 2017, p. 10). Crucial to this approach are Processing Units (PUs) in L2 speech, which can be identified by analysing the phonological coherence, accuracy of form-function mappings, and frequency of learner oral productions.

Research using this approach has shown that whereas for NSs most formulaic sequences impart a processing advantage, for non-native speakers this pertains to transparent and/or very common ones only (Myles & Cordier, 2017). Research on beginner learners in an L2 instructed context (e.g., Myles, Hooper, & Mitchell, 1998; Myles, Mitchell, & Hooper, 1999) seems to suggest that learners rely heavily on formulaic sequences at the beginning of their learning, before being able to break these chunks down as proficiency develops. While advanced learners, on the other hand, seem to recognise which sequences are formulaic for native speakers, they may not process these strings in the same way as native speakers (Boers & Lindstrimberg, 2012).

2.4.4 English as a lingua franca (ELF)

A lingua franca is a language used for communication by various L2 speakers who do not share the same L1 (Mitchell, Myles, & Marsden, 2013). While English is perhaps the most dominant global lingua franca, there are many others in current operation in spheres of trade, tourism, education and other contexts. Around a decade ago, Jenkins (2006) argued that SLA research must begin to consider the widespread growth of the use of English as a Lingua Franca (ELF),

highlighting its importance to notions of interlanguage and fossilisation, and advocating for special conceptual consideration of ELF as a collection of rich language varieties rather than merely *failed* English.

In spite of its relevance, EFL has attracted some criticism. One objection concerns the lack of clarity as to whether ELF should be conceptualised as an emerging variety, an emergent process, a set of linguistic resources, or some combination of these (Sewell, 2013). While early ELF studies (e.g., Seidlhofer, 2001, 2005) sought to identify particular characteristic linguistic features, its proponents (e.g., Cogo, 2011) maintain that ELF is not monolithic or a single variety. Consequently, the field has seen a turn away from specific language forms to pragmatic strategies, processes and practices. ELF is further problematised by considerations related to teaching. From a practical and financial perspective, many learners and teachers may feel uneasy about learning goals that deviate away from what they perceive as 'the language of native speakers' (Sung, 2012, 2013), regardless of the variability, complexities and problems inherent in this concept. Moreover, the promise of social and spatial mobility leave many students reluctant to abandon a target language based on native-speaker norms (Blommaert, 2010; Sewell, 2013).

ELF speakers, like all language users, use metaphor. Conducting important groundwork in this area, Pitzl (2009, 2016) analysed examples of idioms in ELF from the Vienna-Oxford International Corpus of English (VOICE) (Seidlhofer et al., 2013), a 1 million word compilation of naturally occurring, non-scripted face-to-face ELF interactions. In an analysis of a dialogue between L1 Serbian and L1 Maltese interlocutors, Pitzl (2016) records the Serbian's switch to Italian (a language she knew, and that shares many words with Maltese) to convey that many people in Serbia smoke, and that Serbian and Italian languages have an equivalent expression for this: 'fuma come un turco [smoke like a Turk]' (p. 305). From this exchange, Pitzl argued that the L1 Serbian speaker consciously displayed her awareness of the multilingual resource pool, and that "with the key clue *we have a proverb like Italians*…[the Serbian participant] affiliates herself with the speech community of speakers of Serbian (that is, we)" (p. 305).

Both of these claims, however, are problematic. First, how can Pitzl be sure of what the learner is consciously doing? Second, while the Serbian speaker's use of English 'we' might be descriptively categorised as an affiliation with the speech community of speakers of Serbian, it is also possible (and quite likely) that this is simply down to L1 transfer rather than a conscious choice to reposition oneself with one's tribe.[22] Such considerations highlight that while metaphor in ELF is a vastly important and current issue, it is important not to impose one's

[22] In Serbian (and related Balkan languages), the use of 'we/our/ours' to refer to language, people, nation and persons is common, for example, "they [Croats] stole our [Serbian] language" (from research on far-right hate speech, Ilić, 2014, p. 61).

interpretation on the data in an unsupported way.

2.5 Chapter summary

In this chapter, the foundations of research into metaphor in language, thought and communication have been set out. Following that, research into L2 (and L1) metaphoric competence, vocabulary knowledge and language proficiency were critically reviewed. While several findings point to a general relationship between L2 metaphoric competence and L2 vocabulary knowledge and proficiency, studies have contained numerous methodological and other flaws, and leave many important questions such as the conceptual structure of L2 metaphoric competence and extent to which it can be predicted by L2 vocabulary and proficiency measures unexplored. Despite their longstanding contribution to L2 metaphoric competence research, Low (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skills and (sub)competences have never been elicited or used to design metaphoric competence tests. In the next chapter, reliability in metaphoric competence research is examined.

Chapter 3: Literature Review 2

3.1 Introduction

Validity and reliability are important aspects of good research. Cohen, Manion and Morrison (2011, pp. 179-201) summarise that validity has traditionally been understood as the extent to which an instrument measures what it purports to measure, whereas reliability concerns notions of stability, equivalence and internal consistency.[23] Validity can be 'internal', where the extent to which the data warrant and sustain a purported finding or explanation is at stake, and 'external', where the concern is the degree to which the observed results can be generalised from the sample measured to the wider population. Conceptualised as stability, reliability refers to the capability of an instrument to yield similar data from the same (or similar) respondents over time. Equivalence refers to the ability of alternative versions of a data gathering instrument to yield similar results and the agreement between different raters where an instrument requires human judgement. Finally, internal consistency refers to the extent to which a set of items statistically 'hang' together to form a coherent test. One much used measure of internal consistency is Cronbach's alpha (1951), which provides a coefficient (α) of inter-item correlations, namely the correlation of each item with the sum of all other items. While rules-of-thumb tend to stipulate that Cronbach's alpha values of $\alpha > .7$ are acceptable, and those below are not (A Field, 2013) the small study of reliability in metaphoric competence research reported in this chapter (and discussion in subsequent chapters) show that the situation is not so straightforward.

3.2 Reporting and magnitude of reliability estimates in (L1 and L2) metaphoric competence research: A small study

How has reliability been reported in metaphoric competence research? To what extent have metaphoric competence instruments been reliable? Answers to these two questions will lay important groundwork for test development in the present study. Investigations into L2 metaphoric competence form a small part of the wider field of Second Language Acquisition (SLA) research. In SLA, a recent series of meta-analyses (e.g., Plonsky, 2013; Plonsky & Derrick, 2016) have greatly helped researchers to understand the implications and limitations of research findings. These analyses have shown that instrument, interrater and intrarater

[23] Many more (often highly complex) conceptualisations of validity and reliability exist; however, those introduced in this chapter are the most relevant to the present study.

reliability coefficients are not always reported properly, or at all. They have also given empirically grounded benchmarks for reliability, and shown that study, instrument and participant features such as skill type, piloting, and the proficiency of participants all seem to have an effect on reliability estimates. Unsurprisingly, estimates of instrument reliability have been lower than interrater and intrarater reliability, while intrarater reliability estimates have been the highest, but also least reported.

Unfortunately, no equivalent review of reliability in (L1 or L2) metaphoric competence research has been conducted. Conducting a meta-analysis akin to the ones mentioned above is beyond the scope of the present study. Nevertheless, it was important to gain some understanding of the extent to which reliability has been reported in metaphoric competence research, the kinds of estimates used, and what might constitute 'average' levels of instrument, interrater and intrarater reliability. Consequently, a small investigation into these questions involving several of the studies outlined in the literature review was conducted.

3.2.1 Method

Since metaphoric competence research is scattered across various types of publication dating back several decades, use of an overly stringent selection criterion (e.g., *only articles from the top 25 ranking peer-reviewed journals published after 2010*) would likely exclude too much relevant data. Therefore, 33 empirical investigations into metaphoric competence[24] that comprised a list compiled by the author for the present thesis were selected. All appeared in peer-reviewed publications (as articles, books, book chapters, conference proceedings) or approved PhD theses, and were in English. The empirical investigations are, from most recent to oldest: Chen and Lai (2015); Aleshtar and Dowlatabadi (2014); Zhao et al. (2014); Chen, Lin, and Lin (2014); Doiz and Elizari (2013); Taki and Soghady (2013); Beaty and Silvia (2013); Silvia and Beaty (2012); Mashal and Kasirer (2012); Chen (2011); Kathpalia and Carmel (2011); Mashal and Kasirer (2011); NourMohamadi (2010); Chiappe and Chiappe (2007); Azuma (2005); Boerger (2005); Littlemore (2004); Levorato and Cacciari (2002); Littlemore (2001); Boers and Littlemore (2000); Johnson and Rosano (1993); Danesi (1992); Johnson and Pascual-Leone-J. (1989); Gibbs (1980); H. R. Pollio and Smith (1980); H. R. Pollio and Smith (1979); Pickens and Pollio (1979); M. R. Pollio and Pollio (1979b); H. R. Pollio and Burns (1977); Winner and Gardner (1977); Winner et al. (1976); H. Gardner et al. (1975); Steinberg (1970).

Next, a simplified version of Plonsky and Derrick's (2016) coding scheme was developed as a framework for recording study, instrument, participant and substantive features (see Table

[24] An investigation into metaphoric competence was defined as any study that investigated L1 or L2 (L3, L4, etc.) language learners' awareness, retention, comprehension and production of metaphor and other figurative language (i.e., in the same way as in the literature review).

3.1 below). Since all 33 studies were known to be empirical, the first step in the procedure was to identify whether or not instrument, and if applicable[25] interrater and intrarater, reliably estimates had been reported. Following that, data were coded by the researcher. Due to time and funding constraints it was not possible to employ other coders or for the researcher to conduct a second pass of the data, although it would have been advantageous.

3.2.2 Results

In these 33 studies, 176 applications of L1 and L2 metaphoric competence instruments were found, suggesting an average of at least 5 applications per study. When an instrument was administered to a group of participants, this was counted as one application. Any further administrations, even if this involved the exact same instrument and participants, were counted as further applications. Table 3.1 reports the number of applications for which estimates were actually reported. Because all reliability estimates were nonnormally distributed ($p < .01$ for Kolmogorov-Smirnov and Shapiro-Wilk tests) medians and interquartile ranges (IQRs) are used.

Table 3.1 *Reliability Estimates IQRs (176 Instrument Applications in 33 Studies)*

	Number of reports in		Descriptive statistics	
Reliability type	176 applications	33 studies	*Mdn*	*IQR*
Instrument	50	12	.76	.14
Interrater	49	13	.82	.08
Intrarater	0	0	_	_

Note. Only 113 applications required interrater or intrarater estimates.

These data show that instrument reliability was reported in less than a third of 176 test applications, intrarater reliability was reported in less than half of 113 applications in which it could have been, and estimates of intrarater reliability (also reportable in 113 applications) were not provided in any application.

3.2.2.1 Instrument reliability

The first substantive finding was that out of 50 applications in which instrument reliability was reported, 25 of these used Cronbach's alpha. The median α value of .76, shows that on average, estimates were lower in the present study than Plonsky and Derrick's (2016) SLA field median α value of .82, although interquartile ranges in this and the authors' study were similar (.14 and .15 respectively). Table 3.2 (below) shows the proportions and ranges of study, instrument, participant features, and the important finding that estimates ranged from α = .31 to .90 (both Littlemore, 2001). This suggests that even after piloting, the reliability of metaphoric

[25] The use of human raters is not always necessary, for instance when a computer automatically scores reaction times.

competence instruments may be prone to substantial variation.

Given that the number of studies from which data were taken was relatively small (24 compared Plonsky and Derrick's 537), a breakdown of reliability estimates according to study, instrument and participant features was not attempted. Instead, ranges and proportions are reported to identify parameters within which the data lie. Table 3.2 shows that when instrument reliability was reported, most estimates pertain to L1 metaphoric competence, the English language, main study (rather than pilot) data and the receptive mode. Substantial variation can be found in the number of items/tasks (from one to 90),[26] and participants (from six to 149), however, research seems to have focused on pre-teens, teens and young adults (up to undergraduate age) rather than older demographics. L2 metaphoric competence was measured, reliability estimates relate to L1 Japanese, French, Chinese (Mandarin) and Persian learners of English. If non-reports of instrument reliability estimates are considered, the various L1s extend to Italian, Spanish, Cantonese, Sichuanese and Taiwanese Mandarin, Hebrew and Malay and the target languages also include Italian and Spanish.

Table 3.2 *Variation in Instrument Reliability Reported in 50 Instrument Applications*

Feature	Variable	Proportion and/or range
Study		
	L1 or L2 MC	L1 = 28, L2 = 22
	Stage	pilot = 2, main study = 48 (inc. various experiments, pre/post-tests)
Instrument		
	Number of items	1-90 items
	Mode	R = 36, P = 11, not enough information = 3
	Language of	English = 43, French = 7
	Item scoring	Various (e.g., 2- to 5-point scales, reaction times, no. interpretations)
	Scorer(s)	One rater = 17, two raters = 16, three raters = 5, computer = 6, self-report = 6
Participant		
	Number of	6-149 participants (proportion m/f reported in 22 applications)
	Age of	9-21 ('undergraduate' also reported)
	L1	Japanese = 6, French = 12, Mandarin = 8, Persian = 3, English = 21
	TL	English = 27, non-applicable (L1 MC) = 23
	TL proficiency	Various ('low', 'high', 'passed university entrance exam')
Substantive		
	Coefficient[a]	CA = 25, KR = 14, SH = 2, SB = 2, other = 7
	Estimate	0.31 - 0.90 (*M* = 0.75, *SD* = 0.11, *Mdn* = 0.76, *IQR* = 0.14)

[a]TL = target language; CA = Cronbach's alpha; KR = Kuder-Richardson; SB = Spearman Brown; SH = split half.

A final issue, not shown in Table 3.2 is that some researchers (e.g., Aleshtar & Dowlatabadi, 2014) reported reliability estimates from the application of an instrument in another study rather than obtaining estimates from their own sample.

[26] The 'instrument' with one 'item' is found in Silvia and Beaty (2012), who elicited numerous metaphors (each scored on a 5 point scale) via a task that required participants to describe a boring high-school class.

3.2.2.2 Interrater reliability

At .82, the median interrater reliability in this study was also lower than the SLA field median of .92, however the interquartile range of .08 in the present study was less than half that of Plonsky and Derrick's (2016), suggesting less dispersion in the metaphoric competence sample than SLA field more generally. The most common index used was percentage agreement, with Cohen's kappa, and *G* coefficient on a Cronbach's alpha scale used in three applications. In only 26 out of 49 applications were disagreements followed by a final, revised score. These findings suggest two general problems. First, the common use of percentage agreement means that interrater reliability estimates are likely to be inflated through chance agreements. Second, the fact that only around half of the studies report a second, revised statistic, suggests that many disagreements may be left unresolved.

3.2.2.3 Intrarater reliability

Although it was not found in any study, intrarater reliability could have been reported in exactly 113 instrument applications. Of the 2,244 coefficients in 537 studies meta analysed by Plonsky and Derrick (2016), only 40 were intrarater reliability estimates suggesting that this type of reliability is also the least reported in SLA more generally. These had a median of .95 and interquartile range of .06, suggesting that when SLA researchers have measured the extent to which they agree with their own previous decisions, concurrence has been consistently high. In the absence of any data, one can only speculate that high levels of intrarater agreement are likely to hold for metaphoric competence too.

3.2.3 Summary and implications

This small study has revealed that, in line with the SLA field more generally (Plonsky & Derrick, 2016), reliability in metaphoric competence research is generally underreported. Median values for instrument and interrater reliability fall within Plonsky and Derrick's (2016) lower bound acceptability guidelines, suggesting that, on the whole, metaphoric competence instruments and scoring decisions can be considered less reliable than in the SLA field more generally. The instrument reliability estimates of metaphoric competence instruments varied considerably. The reasons that may account for this include the fact that instrument reliability is known to be lower in cases when tests have fewer than 10 items (as several instruments surveyed did), measure psychological constructs, various (rather than single) constructs (A Field, 2013; P. Kline, 1999; Pallant, 2013), or so-called 'broad' rather than 'narrow' constructs (Peters, 2014). In addition, reliability can lower when tests measure certain SLA subdomains or when participants have lower L2 proficiency (Plonsky & Derrick, 2016). Further complexities include inherent weaknesses of Cronbach's alpha compared to item response theory approaches (Bachman &

Palmer, 2010). While a conclusive answer as to why instrument reliability in metaphoric competence research has been so varied requires a separate study in itself, some combination of the factors listed above seems the logical explanation on the available data.

Taken together, these points suggest that if test development in the present study is to improve on past approaches, a more detailed reporting of instrument, interrater and intrarater reliability is necessary. For tests, this is likely to involve calculating and reporting separate reliability estimates for receptive and productive versions of tests, and participants from different L1 backgrounds. For raters, this likely to involve estimating the level of agreement at different scoring stages, before and after discussions about problem items. It will also be necessary to choose an appropriate measure of interrater and intrarater reliability to account for change agreements in the data.

3.3 Research questions for the present study

In the previous chapter, research into metaphor in language, thought and communication, (L1 and L2) metaphoric competence, L2 vocabulary knowledge and proficiency were presented. In this chapter, the results of a small survey of reliability in (L1 and L2) metaphoric competence research were reported. From these two chapters, three general gaps in L2 metaphoric competence research have been exposed.

First, although Low's (1988) and Littlemore and Low's (2006a, 2006b) conceptualisation of L2 metaphoric competence as a broad range of metaphor-related skills and (sub)competences is advocated in many literature reviews, no test to date has measured L2 metaphoric competence as described by the authors. Most, unfortunately, have been limited in scope. Moreover, instrument, interrater and intrarater reliability estimates in (L1 and L2) metaphoric competence research appear to have been lower that those found in the SLA field, but have generally been underreported. The reliability of (L1 and L2) metaphoric competence instruments has been highly variable. These points suggest the need to develop tests to investigate the extent to which Low (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skills and (sub)competences can be reliably elicited and measured. Because the authors also called for more research into how native speakers understand and use metaphor and how this differs from L2 learners, these tests should be administered to both native and non-native English speakers.

Second, although SEM approaches have shown that data from several L1 metaphoric competence tests can be indicative of (or reduced to) more fundamental underlying (sub)constructs, there does not appear to be any research on the underlying structure of L2 metaphoric competence. Furthermore, it is uncertain how the metaphor-related skills and (sub)competences described by Low (1988) and Littlemore and Low's (2006a, 2006b) might

interrelate, overlap, and/or point to more fundamental subcomponents of L2 metaphoric competence; the authors certainly did not test any of this.

Third, studies to date on the relationship between L2 metaphoric competence and L2 vocabulary knowledge (e.g., Azuma, 2005) and language proficiency (e.g., Aleshtar & Dowlatabadi, 2014) have been very limited in number, and have not investigated the extent to which L2 metaphoric competence can be predicted by these measures, or time spent in an L2 immersion setting. In addition, it is uncertain how the strength of relationship between L2 receptive and productive metaphoric competence changes as L2 proficiency increases, and what this might reveal about the development of these skills.

In response to these gaps, six research questions (RQs) were developed. These research questions are grouped into three analyses:

Analysis 1: The development and reliability of the Metaphoric Competence (MC) Test Battery, and descriptive statistics

RQ1: To what extent can (L1 and L2) metaphoric competence be reliably elicited and measured?

RQ2: How do metaphoric competence test scores appear to differ between groups of English NNSs (L1 Chinese) and NSs of English?

Analysis 2: Metaphoric and other (sub)competences uncovered

RQ3: To what extent do factors underlie the observed L2 metaphoric competence, vocabulary knowledge and proficiency test scores for the NNSs? What kind of (sub)competences might these factors represent?

RQ4: To what extent can the same factors be found in the NNS and combined NNS+NS data, and how do the NNSs' and NSs' factor scores differ?

Analysis 3: Relationships between L2 metaphoric competence, vocabulary knowledge, general proficiency, age of starting to learn English and time spent living in the UK

RQ5: To what extent can L2 vocabulary knowledge (size and depth), L2 proficiency (Oxford Online Placement Test and IELTS), age of starting to learn English and time spent living in the UK predict L2 metaphoric competence test scores?

RQ6: To what extent is the relationship between L2 receptive metaphoric competence and L2 productive metaphoric competence different at various L2 proficiency levels?

In the next chapter, I present the methodology, which describes the rationale behind the selection of participants, testing mode, the development of the MC Test Battery, selection of vocabulary knowledge and L2 proficiency measures and the actual method used. The results and

discussion of Analysis 1 are taken up in Chapters 5 and 6, of Analysis 2 in Chapters 7 and 8, and Analysis 3 in Chapters 9 and 10. Finally, Chapter 11 contains the conclusion to the study.

Chapter 4: Methodology and Methods

4.1 Introduction

In the first part of this chapter, the rationales behind using elicited (rather than naturalistic) methods, the written (rather than spoken) mode, and the choice of participants are provided. In the second part, the development of the MC Test Battery is detailed. Following that, the selection of vocabulary knowledge and L2 proficiency measures are described. Finally, the actual data collection methods used are presented.

4.2 General rationales for data collection: Why use…

4.2.1 …elicitation methods?

In Chapter 2, some advantages and disadvantages of both elicitation and naturalistic methods of investigating metaphoric competence were described. As indicated in the research questions, a decision was taken to target Low's (1988) metaphor-related skills and Littlemore and Low's (2006a, 2006b) (sub)competences via elicitation methods rather than in naturalistic data. The main reason for this was due to the need to target and analyse specific metaphors and functions of metaphor described by the authors.

4.2.2 …the written mode?

In line with the majority of L2 metaphoric competence tests to date, it was decided that the MC Test Battery in the present study should be developed to measure the construct in the written mode. One main reason for and advantage of this was to reduce test taker anxiety (Azuma, 2005) which would likely result from spoken elicitation Second, since it was unclear how well the NNSs would be able to handle the MC Test Battery, the written mode was an arguably easier medium. Because MC tests were untimed, test takers had space to think and access more of their linguistic resources than would be possible in speaking. Thus, confounding variables related to spoken performance such as the trade-off between Complexity, Accuracy, Fluency (CAF) and Lexical Richness (Skehan, 2009) were avoided.

4.2.3 …all L1 Chinese non-native speakers of English?

The L2 participants chosen were L1 Chinese learners of English. Given the pervasiveness of metaphor across different languages, there was no in-principle reason for recruiting or rejecting participants from a particular language background. The only stipulation was that sampling allow for as many L2 English participants as possible to be recruited, in order to maximise the statistical power of the various analyses. Because large numbers of L1 Chinese students study at UK

universities (Chapter 1), 112 NNSs of English with this first language were recruited for the main study. The reason for keeping the L1 of the NNSs the same was in order to eliminate the confounding variable of L1 transfer effects. In other words, a mixed L1 sample of NNSs would be problematic because linguistic and conceptual metaphors in different languages do not all correspond equally to those in English (Deignan, Gabrys, & Solska, 1997; Kövecses, 2010), and may thus have different degrees of facilitating or debilitating effect. For recruitment details, see section 4.6.1.

4.2.4 …native speakers of English?

Three reasons can be given for the use of native speakers in the present study. First, understanding more about how native speakers comprehend and produce metaphor is theoretically important, since many L2 learners will seek to emulate nativelike norms (Littlemore & Low, 2006a, 2006b; Low, 1988). While the use of a native speaker 'base' against which non-native speaker knowledge is measured generally runs contrary to the ELF perspective, its practice is fairly commonplace in SLA research (e.g., Foster & Tavakoli, 2009). A second reason was in anticipation of the fact that during the post-study feedback sessions with the NNSs (section 4.6.4), many of the L1 Chinese participants would want to know how their answers would compare with those of the NSs of English. While it was important not to be over prescriptive, having empirical data on this allowed for the research to provide feedback on both the kinds of responses that the NSs and NNSs gave, and areas where the NSs seem to vary. The third reason for using a NS reference group, related to these two points, was in order to be able to identify which areas of L1 metaphoric competence seem to involve more prototypically BLC- or HLC-type tasks (section 2.4.1).

4.3 Development of the MC Test Battery

4.3.1 Selecting metaphor-related skills and (sub)competences to test

The first step in developing the MC Test Battery was to identify all of Low's (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skills and (sub)competences and evaluate them for possible overlaps, theoretical usefulness and measurability. Since both studies discussed metaphoric competence at different levels of specificity, the list of any distinct 'skill', 'ability', '(sub)competence' or 'construct' mentioned quickly exceeded 40! Consequently, a more practical approach was taken. This involved plotting out the headings and subheadings of both publications[27] and using these as the list of possible metaphor-related skills and

[27] Littlemore and Low's (2006a) book was used rather than the journal article.

(sub)competences to develop tests of.

Several of the authors' constructs proved too difficult to test in isolation, impractical, of limited theoretical interest or otherwise unsuitable for inclusion in the MC Test Battery, and so were rejected for consideration from the start. In Low's (1988) study, the ability "to tell when conventional metaphor is being extended idiosyncratically, or a new metaphor is being coined" (p. 130) was theoretically uninteresting as an either/or phenomenon, and too complex to measure as a matter of degree. Being able to "hazard a guess at…[the] speakers' intentions" (p. 130) was unsuitable because it may have seemed like a mind reading task. A test of "knowledge of Vehicles used to describe more than one topic" (p. 131) was developed and administered in the pre-pilot study, but proved difficult to score and time consuming to refine. Measuring "knowledge of mixing of metaphors" (p. 131-132) was theoretically problematic, not least because Low's questionable suggestion that the basic sense of "abate" relates to a storm (cf. MED). A test of "ability to interpret and control 'hedges' [e.g., sort of, kind of, literally]" (p. 133) was developed and administered in the pilot study but later removed because multiple meanings and functions of these words made their operationalisation as test items too difficult. Measuring "awareness of 'socially sensitive' metaphors" (p. 133), though theoretically interesting, was avoided for ethical reasons.

In Littlemore and Low (2006a), sociolinguistic competence (e.g., knowledge of history and behaviour, social organisation, the physical environment) was avoided as a category because it would too heavily engage "world knowledge"[28] (p. 96). Similarly, strategic competence was not targeted because it would introduce a largely non-linguistic dimension into the data. Manipulative functions (illocutionary competence) was avoided because it seemed to concern sensitive issues such as "political rape" (p. 116), and because testing opinion-shifts while controlling for pre-existing political affiliation (e.g., Steen, Reijnierse, & Burgers, 2014) would be problematic. Several aspects of textual competence (e.g., figurative clusters, overarching metaphor and metonymy) were omitted to minimise the amount of reading required of test takers. Concerning lexico-grammatical competence, demonstratives and tense and aspect were unsuitable for developing test items since they presented only a few often interchangeable, closed-class forms. Interchangeability was also a problem for metaphor in modality, where the metaphorical forces, barriers and paths behind forms such as 'must' and 'should' (Sweetser, 1990) cannot be adequately delineated. Grammatical metaphor (and metonymy) was unsuitable because a test of it would require a lot of reading. Finally, a test of metaphor in phraseological patterning overlapped with the one designed to measure Low's (1988) observation that some Vehicles are more acceptable when they employ a particular word class/are exploited conventionally (section 4.3.5.4).

[28] Knowledge of people, places, events, customs and so on rather than language knowledge.

As a result of these decisions, the eventual MC Test Battery contained tests of six out of the 10 of Low's (1988) metaphor-related skills, and three out of four of Littlemore and Low's (2006a, 2006b) (sub)competences.

4.3.2 Creating two versions of the MC Test Battery and splitting participants into group 1 and group 2

In order for MC tests to maximise coverage of constructs, it was decided that wherever possible, both receptive and productive response data for the same metaphors should be obtained. Consequently, for the pilot and main studies, participants were randomly split into two equally (or approximately equally) sized groups (1 and 2), and two versions of the MC Test Battery created. These versions shall be referred to as version 1 and version 2.

The pilot MC Test Battery version 1 was completed by group 1 NNSs (*N* = 5) and NSs (*N* = 2), whereas pilot version 2 was completed by group 2 NNS (*N* = 5) and NSs (*N* = 2). The main study MC Test Battery version 1 was completed by group 1 NNSs (*N* = 56) and NSs (*N* = 15), whereas version 2 was completed by NNSs (*N* = 56) and NSs (*N* = 16). Metaphors used as receptive test items in MC Test Battery version 1 were used as productive items in version 2, and vice-versa. The two versions were the same in all other respects. Full details on the development of test items are reported in section 4.3.5 below. The procedure for administering the different test items is reported in the method (section 4.6). In Chapter 5, a version parity analysis is used to confirm that group 1 and group 2 scores were statistically equivalent, and provide grounds for merging scores for MC Test Battery versions 1 and 2 into one larger dataset for use in further analyses.

4.3.3 Stages of MC Test Battery development: Pre-pilot, pilot and main studies

Stage 1 involved three pre-pilot studies. In the first of these, NNSs of English (L1 Chinese) (*N* = 3) completed the first draft of the MC Test Battery, thus providing information on whether test items would be suitable (at all) for learners at their proficiency (IELTS 6.5 - 7.5). Despite some difficulties, all three participants appeared keen and capable of interpreting the metaphors presented and producing their own when required. In the second pre-pilot study, English NSs (*N* = 2) discussed these NNS responses and the effectiveness of questions. This information was used to develop the scoring criteria, and provide initial information on subjective differences and areas of agreement between raters. In the third pre-pilot study, a second (larger) draft of the MC test was completed by English NSs (*N* = 2). Questions that these test takers appeared to struggle with were deleted, because they would likely present even more of a challenge to NNS participants in subsequent stages.

Stage 2 was the pilot study, in which NNSs of English (L1 Chinese) (*N* = 10) and English NSs (*N* = 4) completed the pilot MC Test Battery version 1 (group 1) and version 2 (group 2) and the size and depth of vocabulary tests. These data allowed for further refinement of the MC Test Battery (versions 1 and 2) based on statistical assessment of easy and difficult questions, underperforming distractors, time taken, and of the types of responses produced. The comments of five participants, three NNSs of English (L1 Chinese) and two English NSs, who completed the MC tests while thinking aloud were used to identify and refine ambiguous instructions and questions, and to gain some insight into how questions would be approached. The vocabulary size and depth tests chosen (section 4.4) were also shown to be suitable for the NNS participants (cf. Azuma, 2005). Specific refinements to questions, instructions, and scoring criteria made during the pre-pilot studies and piloting are mentioned in the report of each test's development (section 4.3).

Stage 3 was the main study, in which NNSs of English (L1 Chinese) (*N* = 112) and English NSs (*N* = 31) completed the final MC Test Battery (versions 1 and 2) and the vocabulary size and depth tests. The NNSs also completed the OOPT and reported their IELTS scores.

An important aside here is that although raters had originally scored all productive responses for both 'meaning quality' (2,1,0) and 'grammatical accuracy' ('1' for *correct*, '0' for *incorrect*), after much consideration, only the 'meaning quality' dimension was used in the analyses presented in this thesis. Because 'meaning quality' and 'grammatical accuracy' are fundamentally different constructs, conflating these scores would render composite test scores ambiguous, making it unclear whether a test taker had gained marks through productive metaphor knowledge or productive grammatical accuracy. Also, if participants had a 'meaning quality' score of '0' (*incorrect*), their 'grammatical accuracy' while executing a metaphor could not be calculated, since they had likely not produced a metaphor or engaged in the skill in question. Finally, scoring 'grammatical accuracy' in binary terms would be somewhat misleading, since not all grammatical errors are equally problematic (e.g., misplaced apostrophe, missing articles, incorrect number agreement, incorrect pronoun). While one could resolve this by devising a more complex partial credit system for scoring grammatical errors, or a larger group of native speaker judges, the problem would then concern the use of time and resources to make sure each of the errors was identified and ranked consistently.

In the next chapter a series of data cleaning analyses and subsequent removals of 'rogue' participants and items are reported. The 'optimal' set of data obtained was then used for further analyses. Although this data cleaning had a methodological function, it is presented as an analysis chapter because the results obtained helped answer the first research question. The NS data were particularly useful for this data cleaning process, both in terms of giving the researcher-designed *best* answers objectivity, and in developing the scoring 'rules' for two of

the MC tests (section 4.3.5.4 and 4.3.5.5).

4.3.4 Selecting reliability indices and developing the scoring protocol

4.3.4.1 Instrument reliability

In Chapter 3, it was shown that although validity and reliability are integral to good research, instrument reliability in the metaphoric competence field has been varied and underreported. It was important to address this in the present study. Since it is the most commonly used measure of internal consistency (Plonsky & Derrick, 2016), and has been used in past metaphoric competence research, Cronbach's alpha (α) was chosen for measuring instrument reliability. Despite offering more comprehensive reliability estimates than similar classical test theory techniques (e.g., split-half), alpha is not perfect and is less informative for tests eliciting a narrow range of scores, or when multidimensionality of skills is present. While item response theory approaches such as Rasch analysis might be used to overcome these issues (Bachman & Palmer, 2010), they are a significant undertaking, peripheral to the focus of the present study, and require more substantial sample sizes, and were thus deemed unsuitable for present purposes. The results of the instrument reliability analysis are presented in the next chapter, and discussed in Chapter 6.

4.3.4.2 Interrater and intrarater reliability

The reliability of scoring decisions for limited production responses was also checked via interrater and intrarater reliability analyses. To allow for direct comparison with other studies, estimates are reported as percentage agreements, the most commonly used index. To make up for the shortcoming that percentage agreements can be skewed by chance agreements, a second index, weighted kappa (J. Cohen, 1968) was also used. Weighted kappa is one of a handful of indices appropriate for use with two coders and ordinal scoring categories (Feng, 2014) and was developed as an ordinal data equivalent of Cohen's kappa. Due to time constraints and the small size of the pilot study sample, formal interrater and interrater reliability checks were conducted for the main study only, and not during piloting. Scoring of limited production questions involved three distinct stages:

Scoring stage 1: First, rater 1 (the author) scored NNS and NS responses to all limited production questions, namely any group 1 or group 2 question requiring a written/typed response (except for one test).[29] Given that both groups' responses needed scoring, this comprised 72 questions from productive tests 4, 5, 6, 7, 8, 9 scored 0-2 (12 per test), and six

[29] Since there was (on the whole) only one *correct* answer to Test 1-Phrasal Verbs-P questions, scoring was clear-cut and did not require corroboration from second or third raters.

questions from receptive test 2, 'Aa' questions scored 0-1. These tests, and the names they were assigned are presented in section 4.3.5. Next, the author recruited a second rater (English NS), who was briefed on the study and trained by completing several practice examples using the scoring criteria and a glossary of key terms (Appendix A). Answers to the practice examples were discussed to encourage a consistent approach to scoring. Rater 2 then scored all limited production responses in the MC Test Battery (versions 1 and 2) which were compared with rater 1's decisions to calculate interrater reliability estimates and identify disagreements. For each disagreement, raters 1 and 2 then reconsidered their original decision, working independently, and knowing only that a disagreement had occurred (i.e., not the other rater's original score). The revised decisions were then compared, and persisting disagreements resolved during face-to-face meetings to arrive at final rater 1 and 2 decisions.

Scoring stage 2: Five months later, a third rater (English NS), was sought and trained in the same way as rater 2. Rater 3 then scored all responses and the author calculated a second set of interrater reliability estimates by comparing rater 3's decisions with rater 1 and 2's final decisions.

Scoring stage 3: Finally, five months after that, the author (rater 1) conducted a second pass, rescoring all responses without reference to his original decisions. Intrarater reliability estimates between the author (rater 1)'s second pass and rater 1 and 2's final decisions were then calculated. The final scoring criteria for limited production questions is also contained in Appendix A. Results of the interrater and intrarater reliability analyses are presented in the next chapter, and discussed in Chapter 6.

4.3.5 The final MC Test Battery

4.3.5.1 Overview

Table 4.1 presents an overview of the MC Test Battery, and lists the names given to tests, constructs tested, their operationalisation, component parts, number of items (*k*), skill and questions type, and scoring used. Test names correspond to key aspects of the constructs tested, and are tagged as either receptive (-R) or productive (-P) tests. To see the final MC Test Battery version 1, completed by the group 1 participants, the reader is referred to Appendix B.[30] Throughout the MC Test Battery, receptive and productive knowledge was conceptualised as 'recognition' and 'recall'. A translation-based conceptualisation was unsuitable because the researcher did not speak Chinese, and because English and Chinese linguistic and conceptual metaphors do not necessarily correspond. Most receptive tests measured 'form recognition' via four option multiple-choice questions. Exceptions include Test 2-Metaphor Layering-R Part A 'a'

[30] Due to limited space, MC Test Battery version 2 (completed by the group 2 participants) is not presented in the Appendices, however its questions can be inferred from section 4.3.5.

questions (measuring 'form recall' via limited production tasks, see Appendix A for scoring criteria) and Part A 'b' questions (measuring 'meaning recall' via multiple-choice tasks), and Test 3-Vehicle Acceptability-R and Test 4-Topic/Vehicle-R (measuring 'meaning recall' via rating scale tasks).

All receptive tests were scored '1' (*correct*) or '0' (*incorrect*). Because all receptive questions used the same scoring scale, composite (i.e., overall) scores for receptive metaphoric competence in the MC Test Battery could be calculated.[31] For multiple-choice questions, *correct* answers are technically *best* rather than *correct*, since other possible answers (not among the options) may exist[32]. In order to decide how many distractors to use, studies on this issue (e.g., Lee & Winke, 2013; Rodriguez, 2005) were consulted, leading to a decision to develop four-option items (i.e., three distractors plus one *correct* answer). The advantage of four-option items over three-option items is a reduced chance of guessing the *correct* answer. The advantage of four-option items over five-option items is that tests take less time to complete. The NNS and NS pre-pilot and pilot participants also vouched for the normality of four-option item format for them.

All productive tests measured 'form recall' via limited production tasks scored either '2' (*correct*), '1' (*partially correct*), and '0' (*incorrect*). This partial credit system was partly inspired by Azuma's (2005) format and refined during the pre-pilot and pilot studies, and in response to initial disagreements between raters 1 and 2 in the main study. In order to maximise motivation, items for all tests in the main study were presented in the order easiest to most difficult using difficulty scores calculated from the NNS pilot data. For multiple-choice questions, all options (i.e., *best* answers and distractors) were automatically randomised for each test taker.
In the remainder of this section, each test's development is presented in detail.

4.3.5.2 Test 1-Phrasal Verbs-R and -P

Test items

In their discussion of metaphor and lexico-grammatical competence, Littlemore and Low (2006a) referred to prepositions and particles as "a traditional and recurring nightmare for all learners of English" (p. 158) and reported several studies arguing for drawing learners' attention to the prototypical (i.e., basic, concrete) sense of prepositions, or teaching them about conceptual metaphors, such as MORE IS UP / LESS IS DOWN to aid comprehension and production. Consequently, Test 1-Phrasel verbs-R and -P were developed to measure test takers' ability to

[31] Statisticians would probably contest that composites computed from items using different scoring scales (e.g., 3-point and 4-point) are problematic because they can falsely equate test takers with very different scoring profiles.
[32] Test 1-Phrasal Verbs-R and –P are exceptions because most questions had only one particle that could possibly fill-the-gap whilst keeping the meaning of the clue the same.

Table 4.1 *MC Test Battery Overview*

Test name	Construct(s) tested	Operationalised as test of ability to:	Part	*K*	Skill type	Question type	Scored
Test 1-Phrasal Verbs-R	Grammatical competence - phrasal verbs (Littlemore & Low, 2006a, pp. 162-166)	recognise metaphorical phrasal verb particles	A	10	Receptive (form recognition)	Multiple-choice	1,0
Test 1-Phrasal Verbs-P		recall metaphorical phrasal verb particles	B	10	Productive (form recall)	Limited production	
Test 2-Metaphor Layering-R	Awareness of multiple layering in metaphors (Low, 1988, p. 134), Ability to construct plausible meanings (Low, 1988, p. 129)	understand the meaning of linguistic metaphors	Aa	6	Receptive (meaning recall)	Limited production	1,0
		recognise the most relevant aspect of meaning for understanding metaphors	Ab	6	Receptive (meaning recognition)	Multiple-choice	
		recognise endings to garden path sentences (fig-lit)	B	6	Receptive (form recognition)	Multiple-choice	
		recognise endings to garden path sentences (fig-fig)	C	6	Receptive (form recognition)	Multiple-choice	
Test 3-Vehicle Acceptability-R	Knowledge of the boundaries of conventional metaphor: knowledge of which features of the vehicle Y can be exploited conventionally and which cannot (Low, 1988, pp. 130-131)	rate the acceptability of semantic exploitations of Vehicles	A	16	Receptive (meaning recognition)	Rating scale	1,0
	Knowledge of the boundaries of conventional metaphor: Knowledge of Vehicle acceptability across different word classes (Low, 1988, p. 131)	rate the acceptability of Vehicles across different word classes	B	12	Receptive (meaning recognition)	Rating scale	
Test 4-Topic/Vehicle-R	Awareness of acceptable Topic and Vehicle combinations (Low, 1988, p. 132)	rate the acceptability of Vehicles as analogies for given Topics	A	6	Receptive (meaning recognition)	Rating scale	1,0
Test 4-Topic/Vehicle-P		produce Vehicles as analogies for a given Topics	B	6	Productive (form recall)	Limited production	2,1,0
Test 5-Topic Transition-R	Textual competence: Marking the edges of a text-Figurative language in topic transition (Littlemore & Low, 2006a, pp. 144-149)	recognise idioms/proverbs/sayings in topic transition	A	6	Receptive (form recognition)	Multiple-choice	1,0
Test 5-Topic Transition-P		produce idioms/proverbs/sayings in topic transition	B	6	Productive (form recall)	Limited production	2,1,0
Test 6-Heuristic-R	Illocutionary (heuristic) functions (Littlemore & Low, 2006a, pp. 126-129)	recognise similes used to perform heuristic functions	A	6	Receptive (form recognition)	Multiple-choice	1,0
Test 6-Heuristic-P		produce similes to perform heuristic functions	B	6	Productive (form recall)	Limited production	2,1,0
Test 7-Feelings-R	Illocutionary (ideational) functions (Littlemore & Low, 2006a, pp. 112-116)	recognise metaphors that convey feelings about information	A	6	Receptive (form recognition)	Multiple-choice	1,0
Test 7-Feelings-P		produce metaphors that convey feelings about information	B	6	Productive (form recall)	Limited production	2,1,0
Test 8-Idiom Extension-R	Illocutionary (imaginative) functions (Littlemore & Low, 2006a, pp. 129-132)	recognise extensions of the literal senses of idioms	A	6	Receptive (form recognition)	Multiple-choice	1,0
Test 8-Idiom Extension-P		produce extensions of the literal senses of idioms	B	6	Productive (form recall)	Limited production	2,1,0
Test 9-Metaphor Continuation-R	Interactive awareness of metaphor (Low, 1988, pp. 134-135)	recognise continuations of metaphor in discourse	A	6	Receptive (form recognition)	Multiple-choice	1,0
Test 9-Metaphor Continuation-P		produce continuations of metaphor in discourse	B	6	Productive (form recall)	Limited production	2,1,0

recognise and recall (metaphorical) phrasal verb particles.

After surveying the literature, five phrasal verb particles frequently associated with conceptual metaphors ('up', 'off', 'out', 'in', 'down') were chosen. Next, using Gardner and Davies' (2007, pp. 358-359) list of Frequency and Coverage of Top 100 Phrasal Verb Lemmas in BNC, four phrasal verb forms for each particle were selected. In order to control for and investigate the potential relationship between form frequency and item difficulty (see Chapters 7 and 8), phrasal verbs were selected from different bands within the top 100 most frequent. For instance, one phrasal verb with 'up' fell within the 1-25 most frequent band, another in the 26-50 band, a third in the 51-75 band and the forth in the 76-100 band. This was implemented for all particles except 'off', whose most frequent phrasal verb 'take off' was ranked 42, resulting in no 1-25 and two 76-100 representatives. Sentences were then developed in which these verbs were used metaphorically, confirmed by locating equivalent metaphors in the VU Amsterdam Metaphor Corpus (VU AMC) online or by applying MIPVU[33] (Steen et al., 2010).

Table 4.2 contains the full list of 20 phrasal verbs used, along with their item number, receptive multiple-choice distractors, raw form frequency, frequency rank in top 100 list, test item sentences, (possible) corresponding conceptual metaphors and references to where these have been studied. Piloting showed that some distractors had not elicited any NNS responses. Consequently, all items except 8 ('take off') and 11 ('go down') had at least one distractor changed before the main study. The pilot version of item 12, 'pick [out] (choose) a new dress for the formal dinner', was amended to '…a new dress from the selection at the store...', to more strongly convey choosing/picking out as the intended meaning for test takers.

Test format and scoring

Part A (receptive) - Ability to recognise metaphorical phrasal verb particles

Participants encountered all 20 phrasal verbs, 10 in the receptive mode and 10 in the productive mode. Both modes used two items from each of the five particle groups. For this test, items were counterbalanced (see section 4.3.2) so that group 1's receptive items, numbers 1-10, were group 2's productive items and vice-versa.

For receptive questions, scored '1' (*correct*) or '0' (*incorrect*), test takers were informed

[33] For example, the metaphor tagging of 'such investments…could bring in…income' (VU AMC) was treated as evidence of the metaphoricity of the present study test item 'with this new job I can bring _[in]_ (earn) enough money'. The application of MIPVU confirmed the metaphoricity of three phrasal verb test items with no equivalents in the VU Amsterdam metaphor corpus: 'get in' (MRW, MED6 = contextual, MED1 = basic), 'get down' (WIDLII, MED2 = contextual, MED4 = basic), 'go down' (MRW, MED2 = contextual, MED1a = basic).

Table 4.2 *Test 1-Phrasal Verbs-R and -P Item Development*

No.	Phrasal verb	Multiple-choice distractors	Freq. raw	Freq. rank	Test item	Conceptual metaphor[a] (and reference[b])
1	pick up	on; off; away	9037	4	Business has been very poor but we expect it to **pick** _____ (improve) again before Christmas.	1 (KS, 1996)
16	hold up	down; on; out	1624	50	Just park here and unload, you won't **hold** _____(block) any traffic at this time of night.	2 (KS, 1996)
9	break up	off; down; away	1286	59	Schools usually **break** _____(stop) for summer in the middle of July.	3 (KS, 1996)
13	move up	on; in; across	477	99	I want to **move** _____(get promoted) to a more senior position in my company next year.	1 (KS, 1996)
8	take off	out; up; on	2163	42	How many days will I need to **take** _____(be absent) from work after my operation?	4 (Y, 2012)
17	get off	over; through; down	1086	66	They'll probably **get** _____ (escape) with a warning this time; but it was a very stupid thing to do.	5 (Y, 2012)
5	put off	down; away; out	742	82	The tickets are too expensive; people might be **put**_____(discouraged) from attending.	6 (K, 2001)
20	come off	up; over; down	518	95	One of the boxers was much stronger, so we knew who would **come** _____ (emerge) worse.	5 (Y, 2012)
19	go out	off; over; away	7688	7	We don't want the campfire to **go**_____(become extinguished), so let's find more wood.	7 (N, 2007)
3	get out	away; over; on	3545	29	If this information **gets** _____ (becomes public), it will be the end of her career as a politician!	8 (N, 2007)
12	pick out	up; on; off	856	75	I'll help you **pick** _____ (choose) a new dress from the selection in the store	8 (N, 2007)
4	give out	away; off; over	532	94	We asked all teachers to **give** _____(distribute) a general reminder to students.	8 (N, 2007)
10	come in	up; over; on	4814	15	There's been an accident. We're still waiting for more news to **come** _____(arrive).	9 (K, 2001)
18	bring in	up; out; over	2505	37	With this new job I can **bring** _____(earn) enough money to pay my daughter's tuition fees.	10 (N, 2007)
6	get in	on; with; out	1127	63	I'll try to **get** _____ (do) an hour of reading before dinner.	11 (K, 2001)
15	put in	on; through; up	810	78	I'm not asking you to **put** _____(contribute) too much time, just one or two hours a week.	12 (K, 2001)
11	go down	off; out; under	4781	16	He spoke really quickly; did you manage to **get** _____ (record) everything he said?	13 (KS, 1996)
7	put down	on; in; across	2873	32	Do we need to **put** _____(record) any other names on the list of invites?	13 (KS, 1996)
14	get down	through; on; over	1538	51	Don't let the quality of your work **go** _____ (decrease)!	1 (KS, 1996)
2	take down	on; in; up	775	81	The police officer who spoke to us wanted to **take** _____ (record) all of our details.	13 (KS, 1996)

[a] 1 = MORE IS UP/LESS IS DOWN; 2 = OBSTRUCTION IS UP; 3 = COMPLETION IS UP; 4 = STOPPING/CANCELLING IS OFF; 5 = DEPARTURE/SEPARATION/ESCAPE IS OFF; 6 = MOVEMENT AWAY FROM A FORMER STATE IS OFF; NON EXISTENCE IS BEING OUT; 8 = EXPOSED/PUBLIC IS OUT; 9 = RECEIVING INFORMATION IS AN OBJECT ENTERING; 10 = POSSESSION IS CONTAINMENT; 11 = THE MIND IS A CONTAINER; 12 CONTRIBUTING TIME IS FILLING A CONTAINER; 13 = WRITTEN OR RECORDED IS DOWN.

[b] KS = Kövecses and Szabó (1996); Y = Yasuda (2012); K = Kurtyka (2001); N = Neagu (2007).

that they would use clues in brackets to choose the 'right' multiple-choice option for completing two word phrasal verbs and given an example of this and explanation (see Appendix B). The verb part of phrasal verbs was bold for test takers. Distractors, designed to lure test takers with low overall scores for this test, were created to be as plausible as possible by evoking related concepts. For instance, for the test item 'schools usually break [up] (stop) for summer', test takers might be distracted by 'out' if they reasoned along the lines of containment or exit, and 'down' or 'off' if they focused on concepts of inactivity and deactivation. While the performance of pilot study items was assessed by examining which distractors failed to lure any test takers, the main study involved a more rigorous distractor analysis.

Part B (productive) - Ability to recall metaphorical phrasal verb particles

Productive (recall) questions, presented in Part B, differed from receptive questions only in that test takers were required to supply a particle rather than choose one from a list. Responses to productive questions were also scored either '1' (*correct*) or '0' (*incorrect*). For items 17 and 20, 'get away' and 'come away' were deemed acceptable alternatives, and were also scored '1' (*correct*). A score of '0' was awarded if no answer was given. To address the problem that some sentences (e.g., item 12) did not technically require the addition of a particle to be grammatical, test takers were instructed to always type an answer.

4.3.5.3 Test 2-Metaphor Layering-R

Test items

Low's (1988) discussions of "awareness of 'multiple layering' in metaphors" (p. 134) and "ability to construct plausible meanings" (p. 129) suggest that an important aspect of L2 metaphoric competence involves handling language with multiple layers of meaning, for instance newspaper headlines or witty comments. The author provided two examples, the first a joke that fencing is "the art of missing the point" (Alexander, 1983 cited in Low, 1988, p. 134), the second an advertisement informing buyers that a car "…leaves the rest standing" (Low, 1988, p. 134), which activates three different meanings concurrently. In order to measure second language learners' ability to understand different layers of meaning, Test 2-Metaphor Layering-R was developed.

An appropriate format for measuring multiple layering was via garden path sentences, a type of utterance where the reader/listener is led to an interpretation that turns out to be incorrect. For instance, the simile at the beginning of 'time flies like an arrow, fruit flies like a banana' primes the reader to incorrectly interpret '…like a banana' as a comparison, rather than a statement of preference. Because studies (e.g., Roberts & Felser, 2011) have shown that even advanced learners struggle to fully reconcile the different layers of meaning in garden path sentences requiring more substantial revisions, it was decided that test items should progress in

difficulty from 'straightforward' metaphors to puns with more complex meaning layering.

Sourcing test items was problematic because many metaphors with multiple layers of meaning (e.g., puns) are offensive. Eventually, 18 suitable items were obtained from searches of newspapers, magazines, television, and the internet. In order not to minimise the role of world knowledge (section 4.3.1), the names of celebrities, places and cultural references were removed. The first six items (Part A) were 'straightforward' linguistic metaphors (e.g., 'the news lifted her spirits'). The next six items (Part B) were a series of puns in the form of garden path sentences requiring a figurative to be reinterpreted literally (e.g., 'in the mirror I looked like a million dollars, green and wrinkled!'). The final six items (Part C), also puns, were garden path sentences requiring a more complex reinterpretation of a figurative, for instance as another figurative (e.g., 'never trust an atom, they make up everything!').

Table 4.3 (below) lists all 18 test items. Bold text signifies the part of the item with different layers of meaning.[34] Underlined text signifies a *best* answer for receptive multiple-choice questions. For Part A, *best* answers and distractors are listed in the same column. For parts B and C, these are in separate columns because the 'punchline' was the *best* answer (see test format and scoring). Other columns list the different meanings of bold words, and the researcher's process of determining metaphoricity of different senses test takers needed to engage with. This process involved three methods: (1) finding an equivalent VU AMC example tagged as a MRW, (2) applying MIPVU, and/or (3) confirming that the item was an MED 'phrase', and thus, likely to have a common, figurative sense different from the sum of literal senses from constituent words.

The reader will notice that not all items neatly conform to the categories established, or follow the same grammatical structure. One exception in Part C concerns the fact that 'shocked' is reinterpreted from a metaphorical sense ('surprised', a VU AMC metaphor) to a non-metaphorical sense ('electrocuted', a non-MRW by MIPVU). However, one may argue the reinterpretation of this item is complex because 'shocked' in the sense of 'electrocuted' requires knowledge of an EFFECT FOR CAUSE metonymy whereby the visible result of electricity entering the body is used as a word for 'electrocuted'. Another exception concerns 'two faced', which MIPVU would probably consider a non-MRW on account of it having only one adjectival sense in the MED, which the OED shows to have pre-dated other senses (e.g., referring to 'leaves'). However, the vivid imagery evoked by 'two faced', the fact that a backstabber does not literally grow a second face, and the further complexity invoked by the suggesting of 'wearing [a face]', mean that this item would be among the most difficult for second language users to interpret, hence its placement in Part C.

[34] Words have been bolded for the reader, but for test takers they appeared normal.

Table 4.3 *Test 2-Metaphor Layering-R Item Development*

No.	Test item	Multiple-choice options/distractors[a]	Meanings of bold words	Researcher's process for checking metaphoricity of: First sense	Researcher's process for checking metaphoricity of: Second sense
Part A: Non garden path figuratives requiring no reinterpretation					
1	The news **lifted** her spirits	the idea of feeling lighter in the chest; the idea of strength involved in lifting; the idea of the sound of straining as something is lifted; the idea of breathing air into the chest	improved' (fig.)	VU AMC	_
2	She treated us **in a cold way**	the idea of not wanting to have contact with cold things; the idea of the appearance of ice and snow; the idea of temperature on a thermometer; the idea of receiving cold food	indifferent' (fig.)	VU AMC	_
3	They will want to get married sometime in the **distant** future	the idea of travelling towards a destination; the idea that people often get tired when they travel; the idea of needing to buy things for a journey; the idea that it is expensive to travel long distances	long time' (fig.)	MIPVU	_
4	He has a **fiery** temper	the idea that fire can be frightening; the idea that burning things smell; the idea that fire requires oxygen to burn; the idea of using fire for cooking	easily annoyed' (fig.)	MIPVU	_
5	The conscience is man's **compass**	the idea of a true and good direction; the idea that a compass can be broken; the idea that west is good and east is bad; the idea of the price of a compass	moral guide' (fig.)	MIPVU	_
6	TV is **chewing gum for the eyes**	the idea that chewing gum does not have much nutritional value; the idea that chewing gum is colourful; the idea of the shape of a piece of chewing gum; the idea of different brands of chewing gum	unfulfilling' (fig.)	MIPVU	_
Part B: Garden path figuratives (underlined) requiring literal reinterpretation					
7	In the mirror I looked like **a million dollars**, green and wrinkled!	a bundle of paper bills; sick and old; wonderful	great' (fig.) & 'cash' (lit.)	MED phrase/MIPVU	_
8	When **everything's coming your way**, you're in the wrong lane!	you could be involved in a car crash; life is great; life is a disaster	going well' (fig') & 'oncoming cars' (lit.)	MED phrase/MIPVU	_
9	No person **goes** before their time, unless the boss leaves early!	only when they are due to leave this world; they must wait until the end of the working day; everyone leaves this world early	dying' (fig.) & 'leaving work' (lit.)	MIPVU	_
10	My wife's currently **carrying** our first child,	his brother can't wait for him to be born; she's pregnant; he's such a little baby, and so light to hold	pregnant with' (fig.) & 'holding' (lit.)	MIPVU	_

	he's eight years old the lazy little thing!				
11	The only thing **moving** about this actor's performance was his wig!	was his body; was his terrible singing; was his incredible acting	emotionally engaging' (fig.) & 'sliding' (lit.)	VU AMC	_
12	The young fighter had a **hungry** look, the kind you get from not eating for a while!	the kind you get when you really want to win; the kind that means you are ready to quit; the kind that expresses your hunger for food	wants to win' (fig.) & 'needs food' (lit.)	MIPVU	_
Part C: Garden path figuratives (underlined) requiring more complex (e.g., figurative) reinterpretation					
13	My local police chief does a talk **on** drugs, you can't understand half of it!	he stepped off them at the end; he completely covered the topic; it was well-structured but a bit boring	about' (fig.) & 'under the influence of' (fig.)	VU AMC	MIPVU
14	When I found out my toaster was not waterproof I was **shocked!**	I was electrocuted; I was physically traumatised; I was surprised	surprised' (fig.) & 'electrocuted' (fig.)	VU AMC	Exception
15	Never trust an atom, they **make up** everything!	they apply cosmetics to everything; they compensate for everything; they constitute everything	lie about' (fig.) & 'comprise' (fig.)	MIPVU	VU AMC
16	If I were **two faced**, I would not be wearing this one!	I would seek medical help to get one removed; I would talk badly about people without them knowing; I would look sad	insincere' (fig.) & 'masks' (fig.)	Exception	MED phrase/MIPVU
17	My friends and I put together a performance on puns; it was basically just **a play** on words!	a manipulation of language; a fun time with grammar; a show about sentences	show' (fig.) & 'joke' (fig.)	MIPVU	MED phrase/MIPVU
18	True friends **stab** you in the front!	stab you in the little finger; stab you in the heart; stab you in the back	hurt you' (fig.) & 'hurt but do not deceive you' (fig.)	MIPVU	MED phrase/MIPVU

[a]Underlined words signify *best* multiple-choice answers. Bold words (not bold for test takers) signify the part of the item with different layers of meaning

Test format and scoring

Part A (receptive) - Ability to (a) understand the meaning of linguistic metaphors and (b) recognise the most relevant aspect of meaning for understanding linguistic metaphors

Because second language learners are much more likely to have to comprehend multiple metaphor layering than produce their own newspaper headlines, puns and so on, Test 2-Metaphor Layering-R was exclusively receptive. In this test, items were not counterbalanced, and so groups 1 and 2 answered the exact same 18 questions (i.e., this test was identical in MC Test Battery versions 1 and 2).

In Part A, test takers were required to (a) explain the meaning of the six linguistic metaphors (limited production), and (b), select the most relevant idea for helping understand the meaning of those metaphors (four option multiple-choice). Test takers were first provided with instructions, an example and explanation, and then required to work through this part of the test. As measures of receptive knowledge, both (a) and (b) questions were scored either '1' (*correct*) or '0' (*incorrect*). For (a) questions, the meanings that test takers' explanations needed to convey were stipulated for scorers (see Appendix A). The researcher-designed *best* answers to (b) questions were confirmed by two native speakers in a pre-pilot study, and a further four native speakers in piloting. Distractors were designed to touch on less relevant aspects of meaning. For instance, the notion of a 'fiery temper' (item 4) has less to do with the smell of burning, combustion processes and practical functionality than it does with the danger and fear that fire invokes. Distractors that did not lure any NNSs in the pilot study were replaced before the main study. This applied to all items except 4.

Parts B and C (receptive) - Ability to recognise endings to garden-path sentences

Parts B and C, while theoretically involving different levels of complexity, were presented to test takers as one continuous test. For these questions, instructions, an example and explanation were first given. Test takers were then presented with the item minus its 'punchline' (e.g., 'No person goes before their time, _____') and required to choose the best option from four "for making the sentence funny or witty". This format, used in the main study, constituted a substantial revision from the piloted format involving pre- and post-punchline meaning questions. For Part B, distractors used the same principle of engaging figurative, literal but less appropriate, and opposite figurative senses of the bold words. For instance, the *best* answer for item 7, 'green and wrinkled' was accompanied by 'wonderful' (figurative sense of 'a million dollars'), 'a bundle of paper bills' (literal but not an effective punchline), and 'sick and old! (opposite of figurative sense of 'a million dollars') as distractor punchlines. For Part C, distractors were developed around various senses of the bold words. For example, distractors for item 13

centred around different meanings of 'on', item 15 different meanings of 'make up', and so on. The format of Part B and C questions used in the main study was substantially different form the pilot version.

4.3.5.4 Test 3–Vehicle Acceptability-R

Test items

The ability to judge the acceptability of Vehicle terms in the target language underpins several of Low's metaphor-related skills; Test 3-Vehicle Acceptability-R aimed to measure this construct. In order to tap into both semantic and grammatical aspects of Vehicle acceptability, 16 questions were designed to measure learners' "ability to rate the acceptability of different exploitations of the Vehicle Y" (Low, 1988, pp. 130-131), and 12 designed to measure sensitivity to "acceptability of Vehicle terms across different word classes" (Low, 1988, p. 131). Overall scores for Test 3-Vehicle Acceptability-R were therefore composites from these two types of question, with both parts given equal weighting. Because they were developed using different principles, test items for parts A and B are reported under separate sections, however, test format and scoring was the same for all 28 questions, and so is reported for both parts of the test together.

Part A (receptive) - Ability to rate the acceptability of semantic exploitations of Vehicles (questions 1-16)

Low provided theoretical motivation for this construct by discussing Lakoff and Johnson's (1980) observation that while the THEORIES ARE BUILDINGS metaphor has several common expressions relating to walls and foundations such as "the theory needs a better framework" (p. 130), it has very few concerning rooms, stairways, or other decorative or interior details. Low postulates several of his own acceptable innovations such as "there's quite an impressive façade" (p. 130), and argues that literary writers often extend metaphors in this way, rather than create completely new Topic/Vehicle combinations.

To test this construct, a series of items with varying degrees of acceptability were needed. Since Low's examples are all native speaker productions, he seems to have had *nativelike* productions and judgements in mind as the target for second language learners. For this reason, the NS group's ratings were used as the standard against which to judge individual (NNS and NS) test takers' ratings. The rationale for this decision is described in test format and scoring (below).

Test items were developed from four widely acknowledged (e.g., Kövecses, 2010) (possible) conceptual metaphors: ANGER IS FIRE /HOT FLUID IN THE BODY, CHANGE IS MOTION, DESIRES ARE FORCES BETWEEN THE DESIRED AND THE DESIRER, and IDEAS ARE CONSTRUCTRED OBJECTS. For each of these metaphors, two Vehicle exploitations (i.e., linguistic metaphors)

deemed by the researcher to be of higher acceptability were identified from the British National Corpus – Brigham Young University (BNC-BYU Davies, 2004-), and two of lower acceptability were devised. Table 4.4 presents these (possible) conceptual and linguistic metaphors, their researcher-designed acceptability and order of presentation to test takers (No.).

Table 4.4 *Test 3-Vehicle Acceptability-R Item Development (Part A)*

No.	Test item	Researcher-designed acceptability
	ANGER IS FIRE /HOT FLUID IN THE BODY	
1	His **blood began to boil** as he started shouting	higher
6	He **couldn't bottle his anger up** anymore so he started shouting	higher
14	He **bubbled** as he began shouting	lower
16	She **turned orange** as she started shouting at him	lower
	CHANGE IS MOTION	
2	He **slipped** into a depression	higher
5	The project is **going ahead** as planned	higher
3	His body **went** fat after a few years	lower
12	Her hair had almost **arrived at** being grey	lower
	IDEAS ARE CONSTRUCTRED OBJECTS	
4	The whole theory **fell apart**	higher
8	The idea **holds up** in principle	higher
10	The theory was **the colour of brick**	lower
13	We **entered the front door** of the plan	lower
	DESIRES ARE FORCES BETWEEN THE DESIRED AND THE DESIRER	
7	It was an **attractive** proposal	higher
9	To her the drunken man was **repulsive**	higher
11	There was a lot of **electricity between** the dog and ball	lower
15	Their similarities **jerked** them together	lower

Eight additional items used in the pilot study (two per conceptual metaphor, one higher, one lower) were eventually cut to reduce the size of the test. Of the three higher and three lower acceptability items per conceptual metaphor used in the pilot version of the test, the item cut was the one with the most diverse 4 NS ratings (i.e., highest standard deviation).

Part B (receptive) - Ability to rate the acceptability of Vehicles across different word classes (questions 17-28)

Low's (1988) examples "the river snaked (its way) through the jungle [higher acceptability]" (p. 131) and "the river was (like/resembled a snake [lower acceptability]" (p. 131) demonstrate that some Vehicle terms are more acceptable when they employ a particular word class. To measure

sensitivity to this, six adjective and six verb metaphors were sourced from the BNC-BYU (Davies, 2004-). For three of each, the word class was altered to form less acceptable Vehicle terms.[35] These items, and their order of presentation ('No.') are listed in Table 4.5.

Table 4.5 *Test 3-Vehicle Acceptability-R Item Development (Part B)*

No.	Test item
	Three adjective metaphors in their ordinary form (higher acceptability):
18	He told a **white** lie
22	He has a **killer** headache
19	She made a **firm** proposal to the client
	Three adjective metaphors altered to another word class (lower acceptability):
17	He **freshened** his ideas (verb)
25	The comment **blunts** (verb)
28	The team are trained to makes calls **coldly**; customers never expect their calls! (adverb)
	Three verb metaphors in their ordinary form (higher acceptability):
20	He tried **to pull the wool over my eyes**
21	He never has time **to shoot the breeze**
24	I **picked up** a job last week
	Three verb metaphors altered to another word class (lower acceptability):
23	We solved the **teased out** problem very easily (adjective)[a]
26	We asked for **a called day** at 6pm (noun phrase)
27	I will give you **a show of the ropes** tomorrow (noun phrase)

[a] In this sentence, 'teased out problem' is akin to 'burned-out car', as is therefore classified as an adjective.

In the pilot study think aloud, one NS expressed uncertainty about whether he should interpret 'coldly' (item 28) to mean *without prior contact*, or *harshly*. Consequently, the clause 'customers never expect their calls!' was added. The reader will notice some minor inconsistencies. For instance, '…pull the wool…' and '…shoot the breeze…' are categorised as verb metaphors (rather than phrases), and the compound noun 'cold call' is treated here as adjective + noun before alteration. In addition, 'cold call' was altered to an adverb, not a verb (like its counterparts). These inconsistencies are attributable to the difficulty of using hard and fast rules to identify and manipulating test items in this way, but are not thought to have made the test any less valid as a measure of the skill Low described.

[35] The BNC-BYU contains examples of all emboldened words used in their sentences with the exception of 'to shoot the breeze', which Macmillan dictionary lists as a phrase. The collocates 'white-lie', 'firm-proposal' and the phrase 'to pull the wool…' were also found in the corpus, whereas 'killer-goal [football]' and 'picked up-work' were adapted to '-headache' and 'job'.

Test format and scoring

For this test, groups 1 and 2 completed the same 28 items. Test takers were informed that 'English native speakers often use expressions which mix ideas and concepts in what seems like quite a strange way', given two acceptable and two unacceptable examples with accompanying explanations and, using a method borrowed from A. Katz, Paivio, Marschark, and Clark (1988), required to rate the acceptability of items (i.e., parts A and B) from 0% (*not acceptable*) to 100% (*perfectly acceptable*). The Vehicle terms were bolded for test takers, in order to focus participants on the part of the sentence they needed to judge.

Because all receptive tests in the MC Test Battery required a dichotomous scoring system (section 4.3.5.1), responses needed to be scored as '1' (*correct*) or '0' (*incorrect*). In order to give the scoring an empirical basis and measure variation in the responses, the NS group's mean ratings and standard deviations were used to establish parameters of correctness. Thus, for each item, any individual (NNS or NS) test taker's rating that fell within the range of one standard deviation above or below the NS mean rating was scored '1' (*correct*), whereas ratings outside this range were scored '0' (*incorrect*). Because this still constituted scoring via stipulated 'rules', rather than norm-referenced scoring (where scores are awarded so as to achieve a bell curve normal distribution), this test, like all others, was criterion-referenced. Data cleaning via the removal of test items with large standard deviations for NS ratings is described in the next chapter.

4.3.5.5 Test 4–Topic/Vehicle-R and -P

Test items

Low suggested that native speakers who innovate to overcome a language's limited resources have "reasonably clear ideas about what Topic and Vehicle combinations would make acceptable and reasonably comprehensible (new) metaphors" (1988, p. 132). Prior associations between words or ease of mental image evoked, he argues, are insufficient factors to account for why some combinations are apt and others not, something which further research must seek to establish. Although his comment on native speaker agreement requires some verification, Low's discussion has strong implications for second language learners, who have smaller vocabulary sizes than native speakers, and whose L2 development is likely to involve a certain amount of experimentation. For these reasons, Test 4-Topic/Vehicle-R and -P was developed to measure test takers' ability to rate the acceptability of Vehicles for a given Topic, and produce suitable Vehicles.

As set out in the literature review, and acknowledged by Low, novel Topic/Vehicle combinations are often more naturally presented as similes, or flagged for tentativeness in some way. One way of measuring this construct while keeping test items 'naturally tentative', was via

the use of analogies, and so a method employed by Tourangeau and Sternberg (1981) was borrowed. While these authors required their participants to rate the aptness of four Vehicle terms within the same domain (e.g., land mammals), and between different domains (e.g., 'technology', 'people', 'sea creatures', 'big cats'), for the present study it was decided that receptive questions should measure the more nuanced skill of rating options within the same domain.

Twelve test items (i.e., analogies) were developed with Vehicle terms from the following domains: BODY (4 items); FOODS (2 items); VISION, TRANSPORTATION HUBS, MATERIALS, OCCUPATIONS, ANIMALS, and MOTOR VEHICLES (all 1 item). Table 4.6 (below) lists items along with their Vehicle domain, receptive options to be rated including researcher-designed *best* (i.e., *most acceptable*) answers.

After the pilot study, the item 5 Topic was changed from 'the cleaners…' to 'the company's internal mail team…' in the hope of making 'blood' a more obvious *best* answer Vehicle. Distractors that were poor at luring NNSs (i.e., attracting high acceptability ratings) were reconsidered before the main study, resulting in at least one change to all items except 5, 6 and 11.

Test format and scoring

Part A (receptive) - Ability to rate the acceptability of Vehicles for given Topics using an analogy framework

For this test, items were counterbalanced so that Group 1's receptive items were group 2's productive items and vice-versa (section 4.3.2). Initially, receptive questions were created using a multiple-choice format, requiring the test takers to simply choose the *best* option from four. However, after focus group comments from two NSs during pre-piloting, it was decided to have test takers rate the acceptability of each of the four options from 0-100% acceptable. A score of '1' (*correct*) was given if a (NNS or NS) test taker's highest rated option was the same as the average (mean) highest rated NS option, *and* if their rating for this option fell within one standard deviation above or below the NS mean rating. Failure to meet either of these two criteria resulted in a score of '0' (*incorrect*). This approach brought a greater degree of objectivity to scoring.

Table 4.6 *Test 4-Topic/Vehicle-R and-P Item Development*

No.	Test item[a]	Vehicle domain	Multiple-choice options (to be rated 0-100% acceptable)	
			best answers	Distractors
1	The CCTV cameras are the ____ of the building.	VISION	eyes	eyeballs; goggles; glasses
2	New products at the end of a long production process are the ____ of large companies	FOODS	fruits	acorns[b]; vegetables; seeds
3	This park is the ____ of our city	BODY	lungs	kidneys; mouth; chest
4	The main argument is the ____ of the essay	FOODS	meat	bread; pasta; rice
5	The company's internal mail team are the ____ of the organisation	BODY	blood	brain; fingers; skin
6	The bee hive is the ____ of the animal kingdom	TRANSPORTATION HUBS	airport	taxi rank; bus station; train station
7	Volcanoes are the ____ of the earth	BODY	pimples	mouths; bruises; blisters
8	Chemical elements are the ____ of life	MATERIALS	building blocks	stones; chains; roof tiles
9	The sales team are the ____ of the organisation	OCCUPATIONS	hunters	shepherds; bakers; farmers
10	Killer whales are the ____ of the sea	ANIMALS	wolves	hyenas; horses; rhinos
11	The outside walls are the ____ of the building	BODY	skin	lips; head; ears
12	Alcohol is the ____ of the drunk person	MOTOR VEHICLES	fuel	steering wheel; trunk/bonnet; engine

[a] Topics are at the start of sentences.

[b] Technically a food in various cultures, but offered to test takers because 'mighty oaks from tiny acorns grow' may be a potential distractor.

Part B (productive) - Ability to produce Vehicles for given Topics using an analogy framework

In Part B, test takers were required to type in an appropriate answer to fill-the-gap. Responses were scored '2' (*correct*) if they formed an analogy that was clearly understandable and made logical sense, '1' (*partially correct*) if the analogy was somewhat understandable and made some logical sense, and '0' (*incorrect*) if the response was not understandable, illogical, more of a literal description than an analogy, or no response was given. The full scoring criteria including illustrative examples and justification, is contained in Appendix A.

4.3.5.6 Test 5-Topic Transition-R and -P

Test items

Littlemore and Low (2006a) convincingly argued for the importance of second language learners acquiring the ability to recognise and produce "idioms, and particularly proverbs and sayings" (p. 144) to summarise the main point of a discussion, offer some overall advice and thus (indirectly) signal that the speaker would like to change topic (Drew & Holt, 1998). Although Littlemore and Low had some success in eliciting idioms in topic transition from L1 Japanese learners of English, insensitive productions such as "it'll iron itself out eventually" (p. 148) in a conversation about an interlocutor's unfaithful girlfriend point to "a serious problem with the 'here's a list of idioms, now have a go at using them' approach" (2006a, pp. 148-149). While this (sub)competence is presented as a feature of spoken discourse, there is no reason why it would not also apply to conversations using the written mode, for instance in online messaging. To measure test takers' ability to recognise and recall idioms, proverbs and sayings in topic transition in interactive discourse, Test 5-Topic Transition was developed.

In accordance with the author's stipulations, 12 idioms, proverbs or sayings that might be used to signify a desire to change conversation topic were identified in MED, OED and BNC-BYU. These are presented in Table 4.7 (below), along with three distractors developed from key words within the target items. Since test items were dialogues of several lines, they are presented in Appendix B and not Table 4.7. While the majority of distractors were obtained from the sources listed above, some were famous quotes (e.g., 'no human being, however great, or powerful, was ever so free as a fish', attributed to English art critic John Ruskin, 1819-1900). Piloting resulted in formatting changes only.

Test format and scoring

Part A (receptive) - Ability to recognise proverbs/idioms in topic transition in interactive discourse

For this test, items were counterbalanced so that group 1's receptive items were group 2's

Table 4.7 *Test 5-Topic Transition Item Development*

No.	Multiple-choice *best* answer (MED phrase)	Keyword	Multiple-choice distractors (based around keyword)
1	there's plenty more fish in the sea	fish	give a man a fish, and you feed him for a day…; no human being...was ever so free as a fish; telling a teenager the facts of life is like giving a fish a bath
2	when in Rome do as the Romans do	Rome	Rome wasn't built in a day; Nero found Rome built of bricks; even the Romans couldn't conquer the blue skies and left it clothed in marble
3	honesty's the best policy	Honesty	better to tell some home truths; the truth is hard to come by; truth is stranger than fiction
4	where there's a will there's a way	will	the spirit is willing but the flesh is weak; never spur a willing horse; you can lead a horse to water but you can't make it drink
5	no use crying over spilt milk	milk	there's milk of human kindness by the quart in every vein; no need to milk it; we're living in the land of milk and honey
6	all's well that ends well	end	all good things must come to an end; the end is nigh; it's the beginning of an end
7	sometimes too many chefs spoil the broth (soup)	cooking	sometimes things go out of the frying pan and into the fire; sometimes it's better to let people stew in their own juices; sometimes it's best to cook up a storm
8	blood is thicker than water	blood	it runs in the blood; blood will have blood; you can't get blood from a stone
9	home is where the heart is[a]	home	there's no place like home; it's great to be home and dry; the lights are on but nobody's home
10	a stitch in time saves nine	stitch	you can't go out if you haven't got a stitch to wear; better not to be stitched up; these things have you in stitches
11	there are three things for sure: taxes, death and trouble[b]	three	best to be three sheets to the wind; third time lucky; two's company, three's a crowd
12	the apple never falls far from the tree[c]	apple	you're like apples and oranges; he's the apple of your eye; an apple a day keeps the doctor away

[a] OED phrase, 4 entries in BNC-BYU.
[b] 'death and taxes' OED quote, 3 entries in BNC-BYU, lyric in 'trouble man' by Marvin Gaye (1972).
[c] OED phrase, 1 entry in BNC-BYU.

productive items and vice-versa (section 4.3.2). For each test item, a short dialogue culminating in the idiom/proverb/saying in question was created. For example, the dialogue to elicit 'there's plenty more fish in the sea!' ran as follows:

> **Speaker A:** Did I tell you that Sarah and I broke up last week?
> **Speaker B:** No! Oh that's so sad, how come?
> **Speaker A:** We just weren't right for each other. I'm so down; I just don't feel like I'll ever meet the right person.
> **Speaker B:** I'm sure you will. I know Sarah was great but don't worry, you know what they say,
> _____

Following Littlemore and Low's (2006a) experience (see above), the discourse between the two speakers was constructed so that the elicited idiom follows three expressions of sympathy and encouragement from speaker B (e.g., 'that's so sad', 'I'm sure you will', 'I know Sarah was great but don't worry'). For receptive questions, scored '1' (*correct*) or '0' (*incorrect*), the instructions informed test takers that 'at the end of a conversation, we often use an expression to summarise the main point, specify some overall advice, and/or let the other speaker know that we would like to change the topic', who were then required to recognise (and select) the best idiom/proverb/saying from four.

Part B (productive) - Ability to produce proverbs/idioms in topic transition in interactive discourse

For productive questions, test takers were instructed to write responses like the ones they had just encountered, and what to aim for or avoid. Responses were scored using partial credit scoring described in Appendix A. The scoring criteria for this test was refined several times. The main problem concerned deciding how to score formulaic, but not necessarily metaphorical productions (e.g., 'like father, like son'). Because Littlemore and Low characterised this skill as involving 'idioms', 'proverbs' or 'sayings', and the task had not specifically requested that learners produce *metaphors*, it was eventually decided to score a production '2' (*correct*) if it finished the dialogue appropriately by way of some proverbial advice or a proverbial summary of the other speaker's situation, '1' (*partially correct*) if it did this somewhat appropriately, and '0' (*incorrect*) if it was illogical, constituted literal advice or a literal summary, or no answer was given.

4.3.5.7 Test 6-Heuristic-R and -P

Test items

Littlemore and Low (2006a) highlighted several ways in which the heuristic functions of metaphor play a central role in education. The authors cite examples of L2 heuristic metaphors from Littlemore's (2005) study, in which EAP students taught each other about their workplaces

by conceptualising the Russian Economic Development Agency's perception of itself as a ray of sunlight (and as a burnt out candle, when perceived by others), the Tanzanian Prime Minister's Office as an elephant, and the Lithuanian Cabinet Office as a spider. Test 6-Heuristic-R and -P aimed to measure test takers' ability to recognise and recall similes to perform heuristic functions.

Despite its effectiveness in Littlemore's (2005) study, her task was unsuitable for present purposes because it engaged intercultural competence, required participants to give a short talk, and relied on their shared expertise in international development.

Instead, because of their simplicity, heuristic similes such as "lava is like sticky treacle…[or] runny butter" appearing in Cameron's (2003, pp. 154-174) study on metaphor in British primary school classrooms were used as a basis for developing test items. While Cameron's study involved metaphor in an L1 context (teacher and pupils), it is quite easy to imagine a scenario in which an L2 speaker such as a teacher, doctor, dentist, or nanny uses a heuristic metaphor or simile to help a child understand something in the world around them.

Twelve entities from the human, natural and physical world were selected. These are presented in Table 4.8 along with *best* answers and distractors for multiple-choice (receptive) questions. In the pre-pilot study, the fact that prompt sentences were presented as 'X is like_____', seemed to lead to both function-based comparisons (e.g., 'the brain is like a computer') and visual comparisons (e.g., 'the brain is like a walnut'). Consequently, where interpretation differences had occurred, 'is' was replaced by another verb (e.g., 'functions [like]', 'behaves [like]', 'sounds [like]'). In the pilot study, neither NS recognised the researcher-designed *best* answer for 'skin functions like____', and so this item was replaced with 'using letters to spell words is like____' for the main study.

Test format and scoring

Part A (receptive) - Ability to recognise similes used to perform heuristic functions

For this test, items were counterbalanced so that Group 1's receptive items were group 2's productive items and vice-versa (section 4.3.2). For receptive questions, scored '1'(*correct*) or '0' (*incorrect*), test takers were informed that the process of explaining concepts, ideas and other things to children often involves comparison, provided with 'good' and 'bad' examples of this, and instructed to select the *best* responses. Distractors were developed using various principles, to be as plausible as possible. For instance, for 'Lava running down the side of a volcano moves like____', the *best* answer 'syrup', and accompany distractors 'jam', 'orange juice', and 'blackcurrant cordial', were all sweet foodstuffs. Whereas for 'An electric current running

Table 4.8 *Test 6-Heuristic-R and -P Item Development*

		Multiple-choice options	
No.	Test item	*best* answer	Distractors
1	The brain works like____.	a computer	a television; a calculator; a (computer) monitor
2	An electric current running through a wire is like____.	water in a pipe	a snake in a pipe; mice in a pipe; peas in a pipe
3	A disease in the body behaves like____.	an army on the attack	a shopper in a shopping mall; a tourist in a city; a transport system
4	Lava running down the side of a volcano moves like____.	syrup	jam; orange juice; blackcurrant cordial
5	Eyelids function like ____.	shutters/blinds	windows; doors; floors
6	Using letters to spell words is like ____.	fitting the pieces of a jigsaw puzzle together	moving pieces in a game of chess; counting pieces of money (coins); eating pieces of food
7	The stomach functions like ____.	a car fuel tank	a car boot (trunk); a car bonnet (hood); a car exhaust
8	The ozone layer functions like ____.	protective bubble wrapping	slices of bread in a sandwich; string wrapped around a present; a polystyrene box
9	The heart functions like____.	a pump	a funnel; a tank; a box
10	The roots of a plant function like____.	a ship's anchors	a ship's oars (paddles); a ship's decks; a ship's cannons
11	Thunder sounds like____.	a hundred horses running	a hundred wolves howling; a hundred cats fighting; a hundred elephants eating
12	Clouds function like____.	bags of water droplets	pools of water droplets; bowls of water droplets; boxes of water droplets

through a wire is like____', the *best* answer 'water in a pipe' was accompanied by distractors that used notions of shape ('a snake…', long and thin, like a pipe), movement ('mice…', running through a pipe), and phrasing ('peas…', which sounds similar to 'peas in a pod').

Part B (productive) - Ability to produce similes to perform heuristic functions

For productive questions, test takers were instructed to type their own answers. A production was scored '2' (*correct*) if the simile formed suitably explained the entity by way of comparison of function, sound, appearance and so on, '1' (*partially correct*) if it did this but with logical problems, and '0' (*incorrect*) if it was not understandable, too literal, or no answer was given (Appendix A).

4.3.5.8 Test 7-Feelings-R and -P

Test items

In their discussion of the ideational functions of metaphor, Littlemore and Low (2006a) argued that improving L2 communicative language ability involves being able to recognise when speakers are using metaphors with affective or evaluative components, and learning how to use metaphor to convey one's standpoint. Test 7-Feelings-R and -P was developed to measure test takers' ability to recognise and produce metaphors that convey feelings about information.

Unfortunately, because the poetry-based activity Littlemore and Low (2006a) suggest for training L2 learners would involve a lot of reading, it was not a good basis for developing a metaphoric competence test in this area. Instead, a shorted test was needed. Although ideational functions of metaphor concern negative evaluations more than positive ones (Littlemore & Low, 2006a; Moon, 1998), a decision was made to keep the balance of positive-negative emotions proportional rather than focus on negative emotions alone. Consequently, twelve feelings[36] (six positive, six negative) was selected. Test items, along with the noun of the feeling involved, its location on the positive/negative (+/-) spectrum, and *best* answers and distractors for receptive multiple-choice questions are presented in Table 4.9. Because this test dealt with quite creative, novel, or otherwise tentative metaphors, items were designed to elicited direct metaphors (i.e., similes). Due to their conventionality,[37] two exceptions, items 4 and 8, sounded unusual as similes ('…one lady who is…like the front runner', '…the project is…like my baby') and so were presented as indirect metaphors (without 'like', 'as', etc.). Piloting resulted in at least one distractor revision per item before the main study.

[36] I use the word 'feelings' as an umbrella term for 'emotions', 'attitudes' and 'standpoints'.

[37] MIPVU would classify 'front runner' (MED spaced and LED hyphenated compound noun with final-word stress) as two lexical units, and code 'front' and 'runner' as MRW. MIPVU would also code 'baby' as an MRW (MED1 = basic, MED4 = contextual).

Table 4.9 *Test 7-Feelings Item Development*

No.	Test item	Feeling noun (+/-)	Multiple-choice options	
			best answer	Distractors
1	Let me tell you about my brother, his bedroom reminds me of a ____.	Annoyance (-)	a rubbish tip	a dustbin; a recycle bin; a wastepaper basket
2	The party was about as interesting as____.	Boredom (-)	watching paint dry	watching the wall get painted; watching paint drip; watching paint crack
3	The choir I heard last night were amazing. Their sound was like____.	Beauty (+)	angels rejoicing	angels praying; angels speaking; angels mourning
4	We've interviewed several applicants so far, but there is one lady who is clearly ____.	Impressiveness (+)	the front runner	the front of the organisation; the front walker; the official front
5	Working with global enterprises would be like ____.	Apprehension (-)	trying to get sheep to sit together	trying to drive past a field of sheep; trying to get sheep to eat; trying to get sheep to make noise
6	My niece is so energetic, she's like a little ____.	Adoration (+)	puppy	bird; beetle; mouse
7	My friend is one of the best sprinters in the country. When she runs at full speed, it's like watching ____.	Thrill (+)	lightning	light; a flame; electricity
8	When I think of...the Smith Project as my favourite. That project is really my____.	Sentimentality (+)	baby	little boy; little one; nephew
9	I was so impressed by the complexity of life of those insects. It was like watching____.	Amazement (+)	miniature civilisations	miniature machines; miniature men; miniature horses
10	At the moment, the players are about as useful as____.	Frustration (-)	an ashtray on a motorbike	a left handed pen; a watch at night; a cigarette during lunch
11	Michelle is about as nice as____.	Aversion (-)	being in the rain without an umbrella	being in the sun with an umbrella; being in the rain with waterproof clothing; being in the sun with sun cream
12	Sandwiches from Nancy's are about as tasty as ___.	Blandness (-)	cardboard	wood; wool; glass

Test format and scoring

Part A (receptive) - Ability to recognise metaphors used to convey information and feelings about that information

For this test, items were counterbalanced so that Group 1's receptive items (1-6) were group 2's productive items and vice-versa (section 4.3.2). For receptive questions, scored '1'(*correct*) or '0' (*incorrect*), test takers were required to choose the *best* option for completing comments to show their feelings to someone they had just met, a stipulation that functioned to discourage de-contextualised responses such as 'the film was as sad as Mike'. Distractors related to *best* answers in a variety of ways. For instance, item 1's *best* answer and distractors were all variants on rubbish disposal sites or containers, whereas item 10's *best* answer ('[like] an ashtray on a motorbike') was accompanied by distractors containing possible (but illogical) depictions of uselessness ('a left handed pen', not a recognised apprentice trick, and 'a watch at night', which might be glow in the dark), and a semantically related option that does not convey uselessness ('a cigarette during lunch').

Part B (productive) - Ability to produce metaphors to convey information and feelings about that information

For productive questions, test takers were instructed to type their own answers and given a 'good' and 'bad' example and explanation. A production was scored '2' (*correct*) if it conveyed the speaker's feelings in a way that would be clearly understandable to a newly acquainted interlocutor, '1' (*partially correct*) if it did this but with problems, and '0' (*incorrect*) if it was not understandable, too literal, or no answer was given (Appendix A).

4.3.5.9 Test 8-Idiom Extension-R and -P

Test items

Learning to produce playful extensions such as "I've been sitting on the fence so long my bottom is beginning to hurt" (Littlemore & Low, 2006a, p. 130) is thought to cause learning gains via destabilisation of the interlanguage system (section 2.2.3.6). While Littlemore and Low (2006a) reported that a few advanced learners were able to productively extend the literal senses of idioms, most found their task difficult and did not write anything. To measure test takers' ability to recognise and produce extensions of the literal senses of idioms, Test 8-Idiom Extension-R and -P was developed.

For the first step of test development, twelve idiomatic MED phrases were identified. Each idiom was then embedded in a sentence eliciting a reference to the idiom's literal sense. For example, 'sit/be on the fence' (original idiom) became '…he's been sitting on the fence so

much that ____'. These items are listed in Table 4.10, along with researched-designed *best* answers and distractors for receptive questions. Because items would be counterbalanced, it was important to keep the tenses of sentences and *best* answers as consistent as possible between items 1-6 and 7-12, so that this would not influence responses. Consequently, two items from the pilot study with tenses that were difficult to harmonise with other items[38] were replaced with 'beat around the bush' and 'taste of [his] own medicine'. Piloting also resulted in at least one distractor revision per item before the main study.

In the final tests, both items 1-6 and 7-12 had at least three sentences using past simple, one using present perfect continuous and one using present simple. Similarly, both sets of items had three *best* answers using past simple, one using past continuous, and one using present perfect simple. The exception was 'break a leg' from items 1-6, which BNC-BYU confirms to be more frequent (i.e., naturally occurring) as an imperative than in past or present simple forms, and which was not matched in items 7-12. A similar consideration led to the decision to have four (out of six) items in each set use prompts culminating in 'that' (e.g., item 1 '…for so long that___').

Test format and scoring

Part A (receptive) - Ability to recognise possible extensions of idioms

For this test, items were counterbalanced so that Group 1's receptive items (1-6) were group 2's productive items and vice-versa (section 4.3.2). For receptive questions, scored '1'(*correct*) or '0' (*incorrect*), test takers were given a short explanation of what an idiom is, told that people often play with or extend idioms to emphasise something or make a joke, and given examples of an idiom in both its original and extended forms. For each receptive question, test takers were required to choose the best option for extending the idiom.

After much consideration, a decision was made to use an inductive approach and have test takers work out from the instructions and example that *best* answers should extend the literal sense. The reason for this was that the alternative, a deductive approach in which test takers are given this information, would transform the task into a 'spot the literal sense' activity, rather than engage the kind of figurative (and creative) thinking described by Littlemore and Low (2006a). This issue is discussed further in Chapter 6.

[38] '…I've bitten off so much more than I can chew, that over the next few weeks…[*best* answer] I'll be digesting day and night' and 'It would be great to kill two birds with one stone. But our problem is that…[*best* answer] the birds are flying miles apart'.

Table 4.10 *Test 8-Idiom Extension Item Development*

No.	Test item	Multiple-choice options	
		best answer	Distractors
1	It's been raining cats and dogs for so long that ____.	we've been forced to call the stray animal collection agency	The street has become flooded; The street has turned into a wildlife park; We've been forced to call the zoo
2	He got such a taste of his own medicine that ____.	He exceeded the recommended daily dosage	He finally understood why everyone was upset with him; He finally understood medical science; he didn't read the label on the back
3	It was a difficult decision. We were so stuck between a rock and a hard place that ____.	Our feet were beginning to resemble fossils	we were getting very worried; our feet were going soft; we were falling into the ground
4	Don't worry, your performance will be great Just go out and break a leg. In fact, go out and____.	come back with crutches	do the very best you can; do something that gets you injured; see where you can break your leg
5	After her email the ball is in my court. But the problem is ____.	I didn't want to play anymore	I wasn't ready to make the next decision; I couldn't hit the ball; I wasn't able to make a proper booking
6	When he said that, he became the first person to really put the problem into words. And he hit the nail on the head so hard that____.	we all felt it go through the wood	he fully explained the problem to us; We saw his head start bleeding; he bought his own hammer
7	Let's cross that bridge when we come to it. Although, since the decision seems likely, let's ____.	start figuring out how to cross safely	prepare to deal with this problem now; prepare to take plenty of pictures of the bridge; call highway maintenance
8	His comment really took the cake. In fact it didn't just take the cake, it ____.	took the whole picnic	was the worst possible thing to say; took a nice piece of cake; was the worst piece of cake
9	He made such a mountain out of a molehill that ____.	he was operating hiking excursions	he was creating stress for everyone; he was creating a walking route; he was looking for a new molehill
10	He beat around the bush for so long that ____.	he got dizzy and fell over	we had to ask him to get to the point; we had to follow him around; he got a full view of the bush
11	She fell so head over heels in love that ____.	she rolled all the way down the hill	she wanted to spend all her time with him; she got lost on the ground; she wanted to buy a new pair of heels
12	He seems to be sitting on the fence about it. In fact, he's been sitting on the fence so much that ____.	his wife has brought him a glass of lemonade and a newspaper	we've become frustrated that he hasn't made a decision; we've asked him when he built his fence; his wife has asked him to get down

To cancel out the possible effect of different pre-existing levels of idiom knowledge, each question presented test takers with the original idiom and its definition, followed by a sentence designed to prime an extension of the idiom. Receptive questions used four option multiple-choice. All twelve researcher-designed *best* answers were endorsed by two NSs in piloting, confirming their objective validity. The reader will observe that the first option in the 'Distractors' column in Table 4.10 extends the common figurative sense of its idiom, while the remaining two extend the literal sense, but in less acceptable ways than the *best* answer. For instance, for item 1, the first distractor refers to a lot of rain (common, figurative sense), whereas the second and third refer to 'wildlife' and 'zoo animals', but these are less appropriate than the *best* answer, since 'cats and dogs' are not typically thought of in these terms.

Part B (productive) - Ability to produce possible extensions of idioms

For productive questions, test takers were given a 'good' and 'bad' example and explanation, and instructed to type their own answers. A production was scored '2' (*correct*) if it extended the literal sense of the idiom in a way that makes logical sense, '1' (*partially correct*) if it did this but with problems, and '0' (*incorrect*) if it was not understandable, extended the common, figurative sense, or no answer was given (Appendix B).

4.3.5.10 Test 9-Metaphor Continuation-R and -P

Test items

The final construct considered for test development was Low's (1988) 'interactive awareness of metaphor'. Although the author did not provide examples, the political interview cited in section 2.2.3.6 is evidence for his assertion that "native speakers are expected to be able to continue a metaphoric discourse coherently once it has started, and presumably to know how to end one when desired" (pp. 134-135). Although Low presented this skill in terms of 'speaker' and 'listener', it is also possible for a metaphor to be continued in a conversation taking place online, via an instant messaging service. To measure test takers' ability to recognise and produce continuations of metaphor in discourse, Test 9-Metaphor Continuation was developed.

In order to develop test items, four scenarios in which a metaphor could be used as 'code' were created. Each scenario contained a dialogue taking place on social media (i.e., online) and three questions eliciting a continued metaphor. In scenario 1, test takers were required to interact with a friend who was announcing her pregnancy through metaphor so as not to alert her children sitting nearby (who might read what was being typed). In scenario 2, test takers corresponded with a friend who was sitting in his workplace and, for fear of nearby colleagues seeing his screen, was using metaphor to report progress on an application for another job. In scenario 3, test takers were required to use metaphor jokingly to chat with a

colleague about a successful third co-worker. In scenario 4, test takers were required to use metaphor, again jokingly, to chat with their (fictional) mother about an active brother who had just dropped by for lunch.

These dialogues are listed in Table 4.11, along with researched-designed *best* answers and distractors for receptive questions. For three scenarios, all questions and *best* answers evoked the same overarching concepts: APPLYING FOR A JOB IS CONDUCTING A SECRET AGENT MISSION (scenario 1), PEOPLE ARE MACHINES (scenario 2), and CONDUCTING OFFICE WORK IS PERFORMING MAGIC (scenario 3). For scenario 1, the dialogue sounded forced with one overarching concept, and so questions and receptive *best* answers used different concepts: PREGNANCY IS BAKING (item 1), SEXES ARE COLOURS (item 2), MORE CHILDREN IS PHYSICAL EXTENSION[39] (item 3).

MIPVU, applied to keywords within the *best* answers revealed that all twelve contained linguistic metaphor. It should be noted that the item 1 *best* answer 'you've got a bun in the oven' was also an MED phrase and that the item 2 *best* answer '…pink or blue' is a recognised (though perhaps outdated) metaphorical symbol for male and female, with three BNC-BYU examples containing this meaning. For item 12, because of the metaphoricity of the preceding dialogue, I departed from MIPVU by treating 'spellbound' (a solid compound) as two lexical units, resulting in 'spell' (n) being coded as a MRW (MED4 = basic, MED3 = contextual) and 'bound' as an WIDLII, since it was uncertain whether to treat this word as a –suffix or past tense of 'bind' (v). Piloting also resulted in at least one distractor revision per item before the main study.

Test format and scoring

Part A (receptive) - Ability to recognise coherent continuations of metaphoric discourse

For this test, items were counterbalanced so that Group 1's receptive items (1-6) were group 2's productive items and vice-versa (see section 4.3.2). For each test, the easier of the two scenarios as shown in the pilot study, was presented to test takers first. The order of items within scenarios needed to be kept the same. For receptive questions, scored '1' (*correct*) or '0' (*incorrect*), test takers were informed that people often have conversations in 'code', in which they talk about one thing as if it were another thing, and given an example of this. They were then told that all the conversations they would encounter took place on social media, and presented with the first

[39] The oldest OED sense of 'extending' concerns forcible straining and physical extension of the body or limbs.

Table 4.11 *Test 9-Metaphor Continuation Item Development*

Scenario	No.	Test item	Multiple-choice options	
			best answer	distractors
1	1	**Mary:** Hey! It's Mary, I've got great news, that I'll tell you in code :)…you know I've been really hungry these past few weeks? Well today the doctor confirmed that I'm eating for two now ;) **You:** Hi Mary, Wow! So you're telling me that ____?	You've got a bun in the oven	you've become one sandwich short of a picnic; you've been baking bread; you've burnt your toast
	2	**Mary:** Yep that's right :D The stork will be paying me a visit around March 15th next year :) **You:** Great! That's fantastic news! What about gender? Will you ____?	be buying pink or blue	be getting it in black and white; be asking for green or red; be wanting yellow or orange
	3	**Mary:** I don't know yet, it's far too early, but I'll be announcing it formally in a couple of weeks. **You:** That's wonderful, I'm so glad to hear that once again, you'll be ____:)	extending the family	holidaying with the family; telling the family; naming the family
2	4	**John:** Hi, it's my lunch break… On my laptop so need to write covertly in case anyone walks past and glances at the screen :)…you remember 'operation C'? **You:** Hi John, haha yes I remember. How ____?	is the operation unfolding	has the operation been organised; has it been to shoot a gun; are the gadgets working
	5	**John:** Well I've been in to assess the lay of the land, me and some rival agents met with a strict panel of drill Sergeants if you know what I mean :) It seems they've chosen their James Bond, yours truly ;) **You:** Wow, that's excellent news! So you are saying ____?	You'll be allied to a different government soon	you'll be going undercover soon; you'll be given a gun soon; you'll be given a car with gadgets soon
	6	**John:** That's right. To be honest, I'm a bit worried about how to switch over from my current operation if you catch my drift :) The crew and captain will not be very pleased that I'm jumping ship! **You:** Well think of it like this: ____. Don't worry, it'll be fine.	every operation comes to an end	every agent loses a few gadgets; every operation costs money; every gadget is useful

3	7	**Your mum:** Jack, the machine called in earlier! **You:** Haha, I know we joke about it, but it's really true; he is a machine! You can always see that he is ____!	switched on and in motion	a car with good safety features; an expensive vehicle; changing the tyres
	8	**Your mum:** You'll never believe it, he steamed over to the house in search of midday fuel, again! **You:** That sounds about right! Even though he left home several years ago, he still comes here for refuel. Why didn't you just tell him ____!?	go back to his own petrol station	ask for diesel fuel; drive more safely; drive on the motorway
	9	**Your mum:** well, it was quite nice to feel like the mechanic again, or at least the petrol station attendant! He actually seemed a bit conked out **You:** Really, well, I'm sure that after receiving his refuelling and a bit of home mechanics, he's now ____!	burning rubber	breaking the speed limit; burning his tyres; breaking his car
4	10	**Peter:** Have you heard, the wizard has done his magic again? I mean the secret magic award **You:** oh yes, I heard Mr magic is due to be ____!	formally recognised for his new and inspiring spells	given a witch's hat; paired up with a witch on a broomstick; put under a spell to weaken his powers
	11	**Peter:** Yes, that's right, his spells have been creating quite a positive stir in the kingdom **You:** Which spell in particular? Will the magic circle commend him for ____?	putting such a spell on our clients	introducing one of our clients to Harry Potter; letting our clients look at his spell books; watching Harry Potter with our clients
	12	**Peter:** I think his main magical achievement was something like that. But he's really all-round enchanting; he's simply been running our show for a long time **You:** I agree, I'm completely ____!	spellbound	spelt out; spell checked; spelt

scenario and asked to choose the *best* response to keep the conversation (and the interlocutor's code) going. Distractors were developed using various principles, to be as plausible as possible. For instance, the *best* answer and distractors for item 9 were designed to be semantically related, and with two 'breaking' and two 'burning options, whereas, item 12 options all contained the minimum lexical item 'spel-', but were not all semantically related.

Part B (productive) - Ability to produce coherent continuations of metaphoric discourse

For productive questions, test takers were informed they would be continuing coded conversations, given a 'good' and 'bad' example and explanation, and instructed to type their own answers. A production was scored '2' (*correct*) if it kept the code going via a metaphor evoking either intended or a different but suitable concept, '1' (*partially correct*) if it did this but with problems, and '0' (*incorrect*) if it was not understandable, not written in code (i.e., literal), or no answer was given (Appendix A).

4.4 Selecting vocabulary knowledge measures

Vocabulary size test

Because the MC battery was a time-consuming measure, it was decided that longer vocabulary size tests such as the VLT (section 2.3.4.1) were unsuitable for present purposes. Instead, a more efficient and user friendly measure of the size of a test taker's vocabulary was sought. The option that met this criterion the best was the YesNo test (Meara & Miralpeix, 2015). Most importantly, VYesNo takes around 10 minutes to complete but has a reportedly more reliable scoring system than its predecessor X_Lex (Meara & Miralpeix, 2015).

Vocabulary depth test

Read's 1998 version of the WAT was selected as the vocabulary depth measure for two reasons. The first reason was on account of Schmitt's (2014) suggestion that the best was of distinguishing vocabulary depth from size is by conceptualising it in terms of a lexical network. Second, using the WAT would allow for direct comparison with other L2 metaphoric competence research (e.g., Azuma, 2005). The WAT has the added advantage that, with 40 stimulus words, it is also a relatively time efficient measure given its coverage, taking around 20-30 minutes.

4.5 Selecting L2 proficiency measures

In order to address research questions 3, 5 and 6, measures of the NNSs' L2 English proficiency were needed. Since participants were all engaged in (or about to commence) studies at UK universities, one option for gathering data on their L2 proficiency was to have them provide IELTS scores. There are several advantages to this. First, by providing scores, participants need

not complete another test, lightening the testing burden for them. Second, as a high-stakes, standardised test, IELTS scores are readily convertible to CEFR levels, which can be used to compare the findings of the present study with other research. However, the use of IELTS scores was problematic because, as reported scores, there would be no way to verify their validity. Some participants may not remember their scores well, for whatever reason, others may report scores above (or below) what they actually attained. For this reason, it was deemed necessary to *measure* participants L2 proficiency. As a robust and efficient measure, administered online and taking on average 30-40 minutes, the OOPT was deemed the most suitable test for this purpose. In summary, data on participants' L2 proficiency were collected in two ways: via their reported IELTS scores, and via their scores on the OOPT, administered by the researcher.

4.6 Method

4.6.1 Participants

The participants in the present study were 112 L1 Chinese NNSs of English (101 females and 11 males) and 31 English NSs (13 females and 18 males).

4.6.1.1 NNSs (L1 Chinese)

Although 112 NNSs completed all tests, 128 had originally started the study but later dropped out. Most NNSs (89%, *n* = 99) were postgraduates[40] enrolled or already engaged in study at UK Universities. The remainder were undergraduates. NNSs were recruited from UK universities via solicitation in classes and by emails disseminated by administrators. Participants were informed that the study focused on 'metaphoric competence', although so as not to prime them with poetic associations, terms such as 'metaphor', 'simile' and 'figurative' were avoided in favour of 'expression' and 'option'.[41] Around 60% (*n* = 67) were based at the University of York, a further 25% (*n* = 28) at the University of Leeds, and the remaining 15% (*n* = 17) at other UK universities including Durham, Sheffield, Reading, East Anglia, Manchester and Loughborough. Most NNSs (94%, *n* = 105) were studying social science degrees, with Education/Applied Linguistics accounting for 45% (*n* = 47) of these. The remainder of NNSs were studying natural science degrees.

The age of NNS participants ranged from 18 to 31 years (*M* = 22.9, *SD* = 2.6) at the time of testing. All participants had learnt English as a foreign language at school in China. The reported age of starting to learn English ranged from 3 to 18 years old (*M* = 9.2, *SD* = 2.7). As

[40] At the point of taking part in the study, 48 participants had already started their course, 60 were engaged in pre-sessional courses, and 4 were due to arrive in the UK within the next few weeks. One participant was a recently graduated PhD student working as a Research Assistant.
[41] Use of the term 'idiom' was unavoidable for Test 8-Idiom Extension-R and –P.

adult L2 speakers studying a range of specialised subjects at higher education institutions, the NNSs were considered to have a 'higher' intellectual profile (Hulstijn, 2012).

4.6.1.2 NSs (L1 English)

A total of 31 NSs completed the MC Test Battery, VYesNo and WAT. Data for one additional NS, who completed part of the MC Test Battery and VYesNo but later dropped out, was not included.

Most NSs (81%, *n* = 25) were currently in or retired from full-time employment,[42] while a minority (19%, *n* = 6) were Postgraduate students at the University of York studying for PhDs in Education, History, and an MSc in Global Marketing. NSs were recruited as a convenience sample relative to the researcher and were all British citizens and first language English speakers based in various parts of the UK. The age of NNS participants ranged from 22 to 68 years (*M* = 39.7, *SD* = 16.3) at the time of testing. Given their age, level of education, and range of occupations, the NSs (like the NNSs) were considered to have a 'higher' intellectual profile (Hulstijn, 2012), and be fairly representative of UK citizens of their socioeconomic status (SES).

4.6.2 Instruments

4.6.2.1 Metaphoric Competence (MC) Test Battery

The MC Test Battery was used to measure metaphoric competence (section 4.3.5, Appendix B).

4.6.2.2 Vocabulary knowledge tests

Size of vocabulary knowledge was measured via the VYesNo test (Meara & Miralpeix, 2015). The 40-item 1998 version of Read's WAT was used to measure depth of vocabulary knowledge.

4.6.2.3 L2 proficiency tests

The NNSs' L2 proficiency was measured via the OOPT which yields overall scores, and those for component Use of English and Listening sections. A second L2 proficiency measure was participants' reported IELTS overall scores, and scores for Reading, Writing, Speaking and Listening sections.

4.6.3 Ethical considerations

The study was not thought to pose severe ethical risks since all test takers were all aged over 18, took part in the study of their own volition and were informed that they could cease

[42] The (formerly/) employed NSs included a Marketing Team Assistant, Metering Engineer, Accountant, Post-doctoral Research Associate, two EFL Tutors, a Leisure Centre Manager, Executive Assistant, Senior Policy Advisor, Solicitor, Mechanical Engineer, Environmental Protection Officer, Trainee paramedic, Infrastructure Project Manager, Adult Education Officer, Nursery Assistant, Salesperson, and eight Retirees with backgrounds in chemical engineering, marine biology, and comprehensive school teaching.

participation at any point without negative consequences for themselves or the researcher. Before commencing data collection, ethical approval for the study was granted by the Ethics Committee of the Department of Education, University of York (where the study took place). All participants read and signed consent forms before taking part in the study (see Appendix C for example of NNS consent form for main study participation). Notwithstanding the low risk, the main ethical issue was potentially causing the NNS and NS test takers anxiety (e.g., feelings of inadequacy), and consequently negatively impacting on the NNSs' studies.

To minimise anxiety, the recruitment email informed potential NNS participants about what the study would involve, how long it would take, that it was not connected to their academic studies, and what these participants could expect in return, namely a £5 cash or Amazon voucher and an invitation to attend a group feedback session to discuss the test and answers. NSs, contacted directly by the researcher, were given the same information but were invited to take part with no formal incentive. All potential participants were invited to ask for further clarity before signing up.

Most participants completed the tests via online links, at home and in their own time, whereas some NNSs attended lab sessions (see section 4.6.4). During the lab sessions NNSs were greeted by the researcher, reminded that participation was unconnected to their studies, and given instructions for taking the tests. Lab participants completed tests in their own time and took an organised 15-minute break half way through the session. The researcher provided refreshments and was present at all times to attend to any problems. Data were kept confidential, and all participants' identities anonymised.

4.6.4 Procedure

Main data collection took place from June to November 2015. The total time needed for NNSs to complete all tests in one sessions with a 15 minute break was around 3 hours, similar to that of other L2 metaphoric competence studies (e.g., Littlemore, 2001). The testing procedure was informed by the need to minimise NNSs' anxiety in taking the metaphoric competence tests (Azuma, 2005), and to attempt the difficult task of recruiting a sufficiently large sample of L2 learners for regression analysis (Plonsky, 2013). Because of these concerns, NNSs were offered the choice of completing tests at pre-arranged lab sessions, or at home in their own time, where they were trusted to work independently, without consulting resources or other people for help. In total, 35 NNSs attended one of four lab sessions consisting of between two and 13 participants, while 77 completed the tests at home, thus allowing the sample size to be increased three-fold. The different test settings had no observable effect on the data, demonstrated by the absence of any statistically significant differences ($p < .01$) between Lab

and 'Home' group test scores[43] and the fact that test setting explained negligible amounts of variance in the L2 MC test scores in hierarchical regression analyses (section 9.5). In addition, NNSs had little incentive to cheat, since all took part voluntarily in order to practice their English, were aware that they would receive detailed post-hoc feedback, and informed that participation was unconnected to their studies. In addition, the fact that the researcher was not present when the 'home' group produced their data is a condition no different to the majority of studies which analyse metaphor use in language corpora (e.g., Nacey, 2013).

NNSs who took tests in the 'Lab' setting first signed the consent form and were given a set of instructions for the three tests. The MC Test Battery was completed first, taking approximately 1 hour 30 minutes. After a 10-minute break, participants proceeded to the VYesNo test (10-15 minutes approximately), WAT (20-30 minutes approx.), and OOPT (30-45 minutes approximately). After completing all tests, NNSs received £5 cash or a £5 Amazon voucher on the spot.

Participants who took the tests at 'Home' (77 NNSs and all 31 NSs) were sent the consent form and links to the MC Test Battery, VYesNo and WAT by email. For NNSs, the OOPT was assigned once participants had started the study. The instructions informed participants to complete the tests as soon as possible, on their own, without consulting dictionaries, the internet or asking friends for help. Ideally participants completed tests on the same day, but were permitted to complete them on different days if necessary. After all data had been collected, NNSs were emailed feedback and invited to a session to discuss the tests and further ways to practice and improve their L2 metaphoric competence. NSs were emailed feedback upon request. All tests were administered online via Qualtrics, apart from the OOPT, which has its own web platform. Data were analysed using SPSS and R. Due to space limitations, data files and R scripts are not presented but are available on request.

In order to minimise the need for the reader to keep flicking back to this chapter, decisions concerning the statistical procedures used are presented in Chapters 5, 7 and 9 rather than in this chapter

4.7 Chapter summary

In this chapter, the rationales for using elicitation methods, the written mode, and the choice of participants were presented. This chapter also covered the development of the MC Test Battery,

[43] In total, 20 x MC, vocabulary knowledge and proficiency tests were checked for 'Home' and 'Lab' group differences using independent-samples *t*-test and Mann-Whitney U Test for normally and nonnormally distributed scores respectively. No differences were significant at the .01 level. 'Home' and 'Lab' group differences found at the $p < .05$ level for the OOPT Use of English section and two MC tests are accounted for by the fact that 94% of the 'Lab' group participants (who scored lower) compared with only 40% of the 'Home' group, were newly arrived students attending pre-sessional language courses to improve their English before starting their degree programmes.

focusing on the selection of skills and (sub)competences to be tested, the development of two equivalent versions, piloting, approaches to reliability, selection of items, test formats and scoring and the selection of vocabulary and L2 proficiency measures. Finally, the actual method used was reported. In the next chapter, results pertaining to the first two research questions are presented.

Chapter 5: Analysis 1 - Development and reliability of the MC Test Battery, descriptive statistics

5.1 Introduction

In the first part of this chapter, results and decisions from a series of analyses to remove outliers in the form of 'rogue' items, participants and tests from the MC Test Battery data are reported. These steps are referred to as 'data cleaning' and, since they help directly answer the first research question, are presented in this chapter as results rather than in the methodology chapter. In the second part of this chapter, descriptive statistics from this refined set of MC Test Battery scores are presented. These help answer research question two. The implications of results, and emerging themes are then discussed in the next chapter.

5.2 Data cleaning

Since the MC Test Battery scores would be used as variables in further analyses (Chapters 7 and 9), the main purpose of data cleaning was to make these data as valid and reliable as possible. Because data cleaning involved producing various indexes of participant and item outliers, distractors, and reliability coefficients, this process also helped show which of Low's (1988) metaphor-related skills and Littlemore and Low's (2006a, 2006b) (sub)competences were most easily and reliably operationalised, elicited and measured.

5.2.1 Creating three separate NNS, NS and NNS+NS data files

The first step was to create three data sets:

1. NNS data file– containing only the NNSs' test responses
2. NS data file – containing only the NSs' test responses
3. NNS+NS data file – containing both the NNSs' and NSs' test responses

Separate NNS and NS data files were needed so that outliers could be identified using cut offs relating to test takers' L1 peers, rather than the whole (NNS and NS) sample. The NNS+NS data file was necessary because variables containing combined NNS and NS data would be needed for the MANOVA in Chapter 7. Up until the instrument reliability analysis in this chapter, the NNS+NS data file was essentially[44] the NNS and NS data sets combined, with participant outliers

[44] One exception was participant 31A, whose scores were removed from Test 5-Topic Transition-R in the NNS+NS data file, but not the NNS data file.

removed, and item outliers for either NNSs or NSs removed for both groups. After the instrument reliability analysis, the NNS, NS and NNS+NS data files retained different 'final' sets of items.

5.2.2 Rating scale outlier analysis

The second step in the data cleaning process was to identify any problematic rating scale items used in Test 3-Vehicle Acceptability-R and Test 4-Topic/Vehicle-R.

For Test 3-Vehicle Acceptability-R, participants rated Vehicle terms from 0-100% acceptable (section 4.3.5.4). Before scoring responses, the NS group ratings needed to be examined to identify items with problematically high standard deviations, implying a lack of NS consensus over acceptability. Since there is no formally established benchmark, the decision on what constituted a problematically high NS group standard deviation was somewhat arbitrary. However, after considering the effect of different cut off points, a decision to delete any of the 28 items with a NS group rating standard deviation of 25 or more[45] was made. This value ensured the best balance between retaining enough items to make the test meaningful and retaining only those for which the NS group showed high levels of acceptability (or unacceptability) agreement. In total, 18 (out of 28) items were retained, 10 designed by the researcher to be of higher acceptability, 8 of lower acceptability. The full list of items retained and the scoring parameters used are shown in Appendix D (Rating scale item outliers: Test 3-Vehicle Acceptability-R).

For Test 4-Topic/Vehicle-R item, which required test takers rate four possible analogy completions from 0-100% acceptable, it was decided that rating scale outliers were items for which the NS group's highest rated option did not match the researcher designed *best* option. Appendix D (Rating scale item outliers: Test 4-Topic/Vehicle-R) lists this information and shows that only item 5 was deleted as a rating scale outlier. Because this test involved novel analogies, it was anticipated that even when NSs agreed on the most acceptable option, variability in ratings would be high for some items. Consequently, an upper limit on the NS group standard deviations was not imposed. This approach led to some unorthodox scoring parameters. For instance, for items 6 and 7 both lower and upper limits fall within 0-100%, meaning that test takers were effectively penalised for giving judgements that were too 'absolute' (i.e., close to the ends of the scale). Although the requirement of tentative ratings may be unsuitable for measuring some constructs, for a test involving novel analogies it was appropriate, even desirable.

[45] For comparison, the standard deviation would be 50.8 in the most extreme case of NS group disagreement (i.e., in which 15 NSs rated an item 100% acceptable and 16 rated it 0% acceptable).

5.2.3 Participant outlier analysis

For the next stage of data cleaning, test takers whose scores lay outside the group mean plus or minus three standard deviations for a particular test were identified as participant outliers. Here, the priority was to detect possible skipping through questions (i.e., not engaging with the skill tested) via unusually low scores, rather than prepare the data for a particular statistical test. Consequently, at this stage, the criterion for deletion was deliberately more liberal than other approaches (cf. Pallant, 2013). For the analyses in Chapter 9, more stringent criteria are implemented to detect univariate and multivariate outliers.

The full list of participant outliers and information needed to interpret the table is given in Appendix E. Once overlaps between the identifications are accounted for, a total of 24 participants (11 NNSs, 13 NSs) were found to be outliers. Most of these relate to MC tests or parts of tests. In all except three cases (discussed below), participants were outliers because they scored less than their group's mean minus three standard deviations for the test or part of test in question, which may indicate skipping questions. Low score participant outliers' responses were removed from all MC and vocabulary tests in question for both the file in question and the NNS+NS data file, but their scores for other tests were allowed to remain. Two NNSs (56A and 12B) who reported low overall IELTS scores of 5.0 were allowed to remain, since their measured OOPT scores (CEFR level B1) were not outliers. In all except one case, if a participant was an outlier in the NNS+NS data file, they were also an outlier in the NNS or NS data file, and so were automatically removed from the NNS+NS data file. One exception to this was participant 31A, whose score for Test 5-Topic Transition was an outlier in the NNS+NS data file but not the NNS data file. Since an extreme score such as this may skew the NNS+NS data in the MANOVA (Chapter 7), this participants score was removed from the NNS+NS data file, but allowed to remain in the NNS data file.

In three cases, participants scored higher than the mean plus three standard deviations. Since these high score outliers relate to IELTs and the OOPT (timed, large scale, standardised measures that are not amenable to cheating) and are not implausibly high, they were believed to be accurate reflections of participants' abilities and were not removed from the data for these tests. While it is possible that 46A and 50A reported their IELTS Writing scores inaccurately, there is no evidence that they did and so these scores were not removed.

In summary, although 24 participant outliers were identified, the scores of only 19 (6 NNSs, 13 NSs) were removed from the data.

5.2.4 Item analysis

Next, an item analysis was conducted on each data file to identify any items that needed removing for being too easy or difficult, or poor at discriminating between high and low ability

test takers. An item's difficulty (index *p*)[46] ranges from '0' (no one answered *correctly*) to '1' (all *correct* answers) (Aiken, 2003). For dichotomous items (scored 0-1), index *p* is calculated as the sum of scores divided by the total number of test takers. For items with partial credit scores (e.g., 0-2) index *p* is the sum of scores divided by two times the number of test takers.

An item's discriminability (index *D*) shows to what extent an item was answered *correctly* by test takers who have a lot of the particular quality that the item is designed to measure, and *incorrectly* by those who have less of that quality (L. Cohen et al., 2011). To calculate index *D*, test takers are ranked according to their total score. Then, the top and bottom 27% of the ranked participants are marked as the higher and lower ability groups respectively. Index *D* is the sum of the higher group minus the sum of the lower group divided by the number of people in one group (for dichotomous items) or the number of people in both groups (for 0-2 partial credit items). The minimum index *D* score is -1, indicating that no higher group and all lower group test takers answered *correctly*. The maximum is +1, indicating that all higher group and no lower group answered *correctly*.

Several considerations went into deciding the optimum index *p* and *D* values and acceptable cut off points. First, indexes *p* and *D* take on different significance in norm and criterion-referenced testing (L. Cohen et al., 2011). For criterion referenced tests like those in the present study, L. Cohen et al. (2011) advise that developing items to differentiate test takers is less important (per se) than for norm referenced tests. Aiken (2003) argues that the optimum item difficulty for this type of test is .50, whereas Thomson and Levitov (1985) note that test reliability for a four item multiple-choice question is highest when index *p* is around 0.625 (i.e., half way between the value expected from pure guessing, 0.25, and the maximum value 1.00). Second, as index *p* becomes increasingly higher or lower than its optimum value of 0.50, the researcher is forced to accept *D* values of less than 0.30 (Aiken, 2003). Third, while it would be reasonable to expect a wide range of index *p* values for items in the NNS and NNS+NS data files,[47] values should (usually) be much higher in the NS data file, resulting in lower index *D* values.

For items in the NNS and NNS+NS data files, a decision was made to take index *p* = .50 as an optimum value (Aiken, 2003) and delete any items with index *p* values between 0.33 and 0.67, and index *D* values of less than .30 (L. Cohen et al., 2011). In line with Aiken (2003), items with index *p* values outside of this range could be retained regardless of their index *D* values. Since they did not discriminate well, the main function of these (i.e., very easy or difficult) items was to motivate test takers, and provide information about metaphors and metaphor-related skills that large numbers of test takers have either mastered, or failed to master.

[46] The letter "*p*" here is unrelated to (and not to be confused with) *p* values used in significance testing.
[47] Since the NNS+NS data file was comprised of 78.3% NNSs and 21.7% NSs, the range of p values would be more similar to the NNS data file than the NS data file.

The index *p* values for items in the NS data file were expected to be higher, making the optimum value and cut off parameters just mentioned inappropriate. Although the goal of criterion referenced tests is not to differentiate high and low ability test takers per se, it was important that the items retained allowed for *some* differentiation. Consequently, a decision was made to delete NS items with (1) index *p* values of less than 0.50,[48] regardless of their index *D* values, (2) items with index *p* values from 0.50 to 0.67 (cf. L. Cohen et al., 2011 criteria above) if their index *D* value was less than 0.30, and (3) items with higher NNS index *p* values than NS index *p* values (deleted in all files). Table 5.1 summarises these decisions:

Table 5.1 *Item Analysis Criteria for Removing Items*

Data file	Type of analysis	Optimum value	Criteria for unacceptable item	Based on
NNS	Item difficulty	0.50	Index *p* within 0.33 to 0.67, index *D* below 0.30	Cohen et al. (2011), Aiken (2003)
	Item discriminability	1.00	Index *p* within 0.33 to 0.67, index *D* below 0.30	Cohen et al. (2011), Aiken (2003)
NS	Item difficulty	.50 to 1.00	Index *p* below 0.50	Cohen et al. (2011), Aiken (2003), Thomspon & Levitov (1985)
	Item discriminability	1.00	Index *D* below 0.30 if index *p* within 0.50 to 0.67	Cohen et al. (2011), Aiken (2003), Thomspon & Levitov (1985)
NNS+NS	Item difficulty	0.50	Index *p* within 0.33 to 0.67, index *D* below 0.30	Cohen et al. (2011), Aiken (2003)
	Item discriminability	1.00	Index *p* within 0.33 to 0.67, index *D* below 0.30	Cohen et al. (2011), Aiken (2003)
NNS, NS	NNS and NS comparison	_	NNS index *p* higher than NS index *p*	_

Tables 5.2 and 5.3 (below) list the 'rogue' items that were identified and deleted from the data, and the reason for deletion. Once overlaps in identification are accounted for, the tables show that 19 items were removed from the data, 7 from the NNS data file only, 8 from the NS data file only, and 4 from both files.

[48] For four option multiple-choice receptive questions, this cut off is above *p* = .25, the value expected for pure guessing. For productive questions scored 0, 1 or 2, *p* = .50 (the lowest value permitted) would imply that either one third of the NS test takers achieved each of the three possible scores, or half of the test takers scored 2 and the other half scored 0.

Table 5.2 *Item Analysis List of Rogue Items*

Test	Group	Item	Item content	Data file	Diff. (*p*)	Discr.(*D*)	Problem
Test 1-Phrasal Verbs-R	2	19	go out [choose particle]	NNS	0.38	-0.06	*p*>0.33<0.67 and *D*<0.30
Test 1-Phrasal Verbs-R	2	19	go out [choose particle]	NNS	0.38	-0.06	NNS negative D
Test 1-Phrasal Verbs-P	2	9	break up [produce particle]	NNS	0.36	0.25	*p*>0.33<0.67 and *D*<0.30
Test 2-Metaphor Layering-R	1 & 2	2b	a cold way [relevant feature]	NNS	0.52	0.20	*p*>0.33<0.67 and *D*<0.30
Test 2-Metaphor Layering-R	1 & 2	3b	distant future [relevant feature]	NNS	0.67	0.17	*p*>0.33<0.67 and *D*<0.30
Test 2-Metaphor Layering-R	1 & 2	6a	chewing gum for the eyes [explain]	NS	0.48	0.63	NS *p*<0.50
Test 2-Metaphor Layering-R	1 & 2	18	stab you in the front [best ending]	NNS	0.33	0.23	*p*>0.33<0.67 and *D*<0.30
Test 2-Metaphor Layering-R	1 & 2	9	before their time [best ending]	NS	0.97	-0.13	NS negative *D*
Test 2-Metaphor Layering-R	1 & 2	2b	a cold way [relevant feature]	NS	0.66	0.25	*p*>0.50<0.67 and *D*<0.30
Test 2-Metaphor Layering-R	1 & 2	13	a talk on drugs [best ending]	NS	0.45	0.13	NS *p*<0.50
Test 2-Metaphor Layering-R	1 & 2	2b	a cold way [relevant feature]	NNS+NS	0.55	0.26	*p*>0.33<0.67 and *D*<0.30
Test 2-Metaphor Layering-R	1 & 2	16	two faced [best ending]	NNS	0.59	0.27	*p*>0.33<0.67 and *D*<0.30
Test 2-Metaphor Layering-R	1 & 2	18	stab you in the front [best ending]	NNS+NS	0.38	0.26	*p*>0.33<0.67 and *D*<0.30
Test 4-Topic/Vehicle-R	1	6	bee hive [rank & range]	NS	0.33	0.40	NS *p*<0.50
Test 4-Topic/Vehicle-R	2	7	volcanoes [rank & rate]	NS	0.19	0.20	NS *p*<0.50
Test 4-Topic/Vehicle-P	2	6	bee hive [produce Vehicle]	NS	0.25	0.60	NS *p*<0.50
Test 6-Heuristic-P	1	10	the roots of a plant [produce simile]	NNS	0.38	0.25	*p*>0.33<0.67 and *D*<0.30
Test 6-Heuristic-P	2	3	a disease [produce simile]	NS	0.53	0.20	*p*>0.50<0.67 and *D*<0.30
Test 8-Idiom Extension-R	1	1	raining cats and dogs [choose idiom ext.]	NS	0.40	0.80	NS *p*<0.50
Test 9-Metaphor continuation-R	1	5	so you are saying…?[choose job metaphor]	NS	0.27	0.60	NS *p*<0.50

Table 5.3 *Rogue Items Identified by Comparison of NNS and NS Item Difficulty Indexes (p)*

Test	Group	Item	Item content	NNS *p*	NS *p*	Problem
Test 4-Topic/Vehicle-R	2	10	killer whales [rank & range]	0.79	0.50	NNS *p* > NS *p*
Test 4-Topic/Vehicle-P	2	4	main argument [produce Vehicle]	0.64	0.63	NNS *p* > NS *p*
Test 4-Topic/Vehicle-P	2	6	bee hive [produce Vehicle]	0.27	0.25	NNS *p* > NS *p*
Test 4-Topic/Vehicle-P	1	10	killer whales [produce Vehicle]	0.80	0.60	NNS *p* > NS *p*
Test 4-Topic/Vehicle-P	1	12	alcohol [produce Vehicle]	0.62	0.53	NNS *p* > NS *p*
Test 6-Heuristic-P	2	2	an electric current [produce simile]	0.58	0.53	NNS *p* > NS *p*
Test 6-Heuristic-P	2	3	a disease [produce simile]	0.54	0.53	NNS *p* > NS *p*

5.2.5 Distractor analysis

A distractor analysis was conducted to determine the effectiveness of multiple-choice options designed to lure lower ability test takers from the *correct* response. This analysis was primarily conducted to help evaluate the final MC Test Battery, and to provide data for comparing NNS and NS response patterns (see Chapter 7). Thus, it was not used to delete test items.

For distractors, utility scores range from +1 to -1. A distractor utility score of '1' indicates that the distractor performed perfectly because all lower and no higher ability test takers selected it, '0' that the same number of higher and lower ability test takers selected it, and '-1' that all higher than no lower ability test takers selected it, the worst case scenario. Table 5.4 presents, for both NNS and NS groups, the mean and median distractor utility scores for each receptive MC test involving multiple-choice, the MC Test Battery-R, and the WAT.

Table 5.4 *Distractor Analysis Utility Scores (Descriptive Statistics)*

		NNSs (*N* = 112)					NSs (*N* = 31)				
Test	*K*[a]	*M*	*SD*	*Mdn*	*IQR*	Rk*(M)*	*M*	*SD*	*Mdn*	*IQR*	Rk*(M)*
T1-Phrasal Verbs-R	60	0.63	0.61	1	1	5	0.03	0.18	0	0	6
T2-Metaphor Layering-R	54	0.59	0.74	1	0.75	6	0.19	0.52	0	0	3
T4-Topic/Vehicle-R	36	0.33	0.96	1	2	8	-0.11	0.98	-1	2	7
T5-Topic Transition-R	36	0.72	0.45	1	1	4	0.08	0.28	0	0	5
T6-Heuristic-R	36	0.83	0.51	1	0	1	0.17	0.38	0	0	4
T7-Feelings-R	36	0.83	0.38	1	0	1	0.17	0.38	0	0	4
T8-Idiom Extension-R	36	0.39	0.87	1	1.25	7	0.39	0.60	0	1	1
T9-Metaphor Cont.-R	36	0.81	0.47	1	0	3	0.25	0.44	0	0.25	2
MC Test Battery-R[b]	330	0.64	0.20	1	0	_	0.15	0.15	0	0	_
Word Associates Test	160	0.85	0.49	1	0	_	0.54	0.56	1	1	_

Note. K = number of distractors, Rk = rank.

[a] For all tests except T2, 50% of *K* distractors encountered by group 1, 50% by group 2.

[b] *M, SD*, *Mdn* and *IQR* of all receptive MC test statistics.

Figures 5.1 and 5.2 show box-and-whisker plots of the data to allow for easier comparison of distractor utility for NNS and NS groups.[49] The results show that distractors, on the whole, were better at luring low ability NNSs than low ability NSs. The MC Test Battery-R statistics show that distractors had a mean utility score of 0.64 (0.20) for the NNSs, and 0.15 (0.15) for the NSs. For the NNSs, distractors performed best for Test 6-Heuristic-R, Test 7-Feelings-R, Test 9-Metaphor Continuation-R and the WAT, and worst for Test 4-Topic/Vehicle-R and Test 8-Idiom Extension-R. For the NSs, distractors performed best for the WAT, Test 8-Idiom Extension-R and Test 9-Metaphor Continuation-R, and worst for Test 4-Topic/Vehicle-R, Test 1-Phrasal Verbs-R and Test 5-Topic Transition-R. For both NNSs and NSs, the test for which the performance of distractors

[49] Mean = crosses within boxes; Median = horizontal lines within or on (short) edge of boxes; IQR = boxes; Q1 and Q3 = bottom and top (short) edges of boxes; outliers (> Q3 + 1.5 times IQR or < Q1 – 1.5 times IQR) = points.

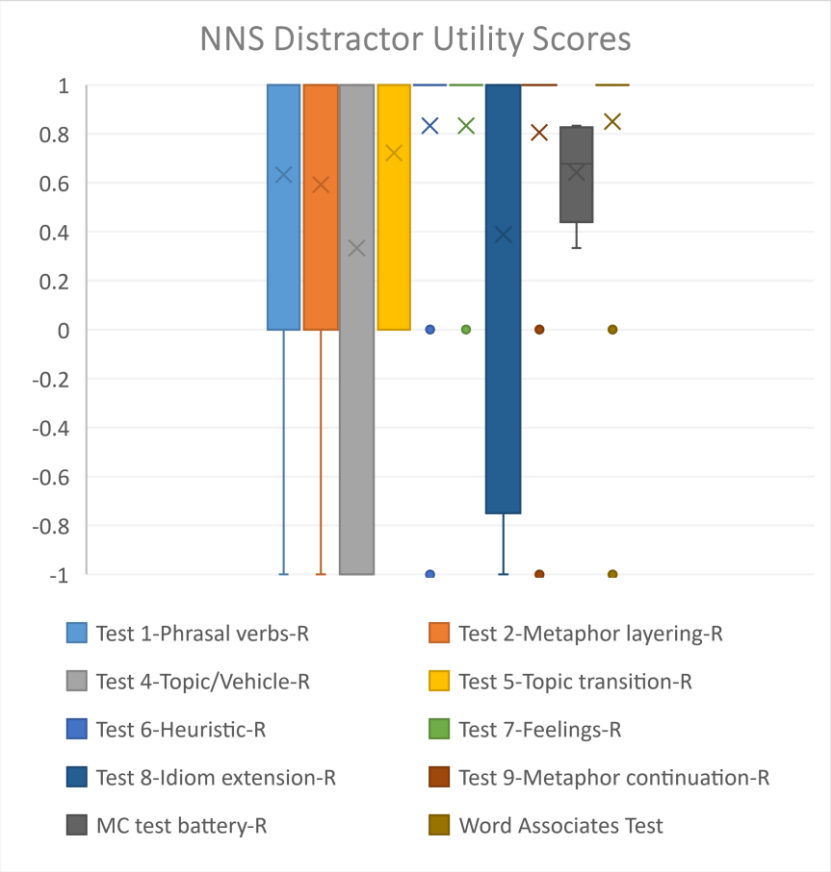


Figure 5.1 NNS distractor utility scores

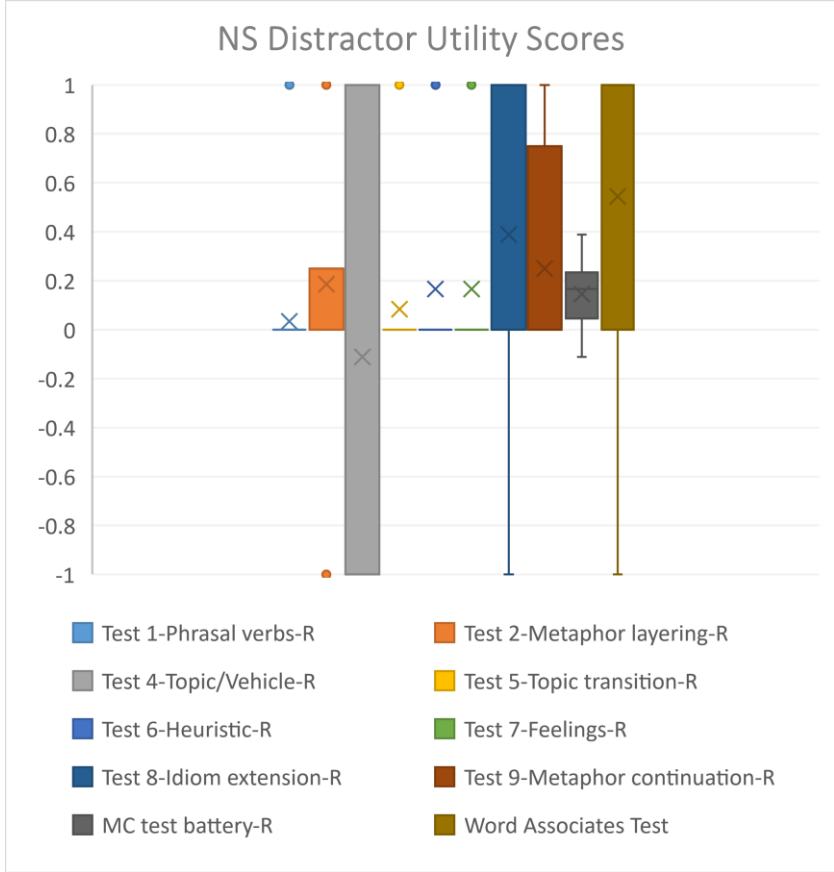


Figure 5.2 NS distractor utility scores

was most varied was Test 4-Topic/Vehicle-R.

5.2.6 Instrument reliability analysis

For the next stage of data cleaning, an instrument reliability analysis was conducted within each data file on the MC Test Battery items retained thus far and the WAT test items.[50] The instrument reliability analysis had three aims:

1. To retain the most internally consistent set of items (i.e., those with the highest alphas)
2. To retain group 1 and 2 items with no statistically significant differences to allow for test scores to be merged into larger, single variables (section 5.2.8)
3. To retain as many items as possible whilst adhering to the first two aims

To achieve these aims, solutions to a number of procedural issues were needed. Since the reliability coefficient could be increased incrementally via the deletion of certain items, there was some tension between the first and third aims. A decision was made to stop deletion at four items remaining, or when the alpha value reached Plonsky and Derrick's (2016) lower acceptable threshold of .74, whichever occurred first. Because group 1 and 2 data were later merged (section 5.2.8) four items remaining meant a set of eventual test scores from eight different exemplars.[51] Concerning the second aim, for a minority of tests (indicated by [a] in Table 5.5), in order to ensure statistical equivalence between group 1 and 2 versions, the items eventually retained were not the most internally consistent.

In the NS data, some alphas were very low, incalculable or negative, due to zero or negative variance caused by all test takers achieving perfect or near perfect scores. While this was expected and theoretically valid (Haladyna, 2004), it made the use of the SPSS option 'scale if item deleted' to remove items logistically problematic, because selecting this option automatically deletes all items with zero variance, often resulting in fewer than four items remaining. For these tests, indicated by [b] and [c] in Table 5.5, the 'scale if item deleted' function was not used, and all items that had not been previously deleted during the item analysis were retained.

[50] Item by item data (required for internal consistency analysis) was available for these tests only.
[51] In cases where two different sets of items are completed by two different sets of participants, it is not possible to compute one Cronbach's alpha coefficient for both sets of items. Rather, two coefficients (one for each set of items) are available.

Table 5.5 *Instrument Reliability of MC Test Battery and WAT: Items-within-Tests*

			NNS data file				NS data file				NNS+NS data file			
Test	R/P	Group	*N*	*K*1[a]	*K*2	α	*N*	*K*1[a]	*K*2	α	*N*	*K*1[a]	*K*2	α
Test 1-Phrasal Verbs	R	1	56	10	4	.33	14	10	10[c]	_	70	10	9[b]	.72
		2	56	9	7	.50	16	10	10[c]	-.14	72	9	8	.74
	P	2	56	9	4	.51	15	10	10[c]	.43	71	9	8	.69
		1	56	10	4[b]	.10	15	10	10[c]	-.19	71	10	7[b]	.73
Test 2-Metaphor Layering	R	1&2	111	20	11	.59	29	20	6	.76	140	17	16	.74
Test 3-Vehicle Acceptability	R	1&2	112	18	10	.72	30	18	11	.54	142	18	18	.88
Test 4-Topic/Vehicle	R	1	55	6	4	.39	15	4	4	.36	70	4	4	.44
		2	56	5	4	.40	16	4	4	.52	72	4	4	.47
	P	2	56	4	4	.33	16	4	4	.83	72	4	4	.50
		1	56	4	4	.26	15	4	4	.21	71	4	4	.33
Test 5-Topic Transition	R	1	56	6	4	.33	14	6	6[c]	_	69	6	4	.36
		2	55	6	4	.46	16	6	6[c]	-.12	71	6	4	.52
	P	2	56	6	6	.62	16	6	4	.39	72	6	6	.76
		1	56	6	4	.60	14	6	4	.40	70	6	6	.75
Test 6-Heuristic	R	1	56	6	5	.50	15	6	6[d]	-.09	71	6	6	.61
		2	56	6	5	.45	15	6	6[c]	.25	71	6	4	.54
	P	2	56	4	4	.45	16	4	4	-.11	72	4	4	.52
		1	56	5	4	.52	15	6	4	.44	71	5	5	.57
Test 7-Feelings	R	1	56	6	5[b]	.46	15	6	6[c]	-.20	71	6	5[b]	.54
		2	56	6	6[b]	.09	15	6	6[c]	.32	71	6	6[b]	.50
	P	2	56	6	4	.51	16	6	4	.50	72	6	5	.64
		1	56	6	4	.33	15	6	4	.38	71	6	4	.50
Test 8-Idiom Extension	R	1	56	6	4	.73	15	5	5	.72	71	5	5	.72
		2	56	6	4	.56	16	6	4	.63	72	6	5	.76
	P	2	56	6	6	.84	16	6	6	-.24	72	6	6	.89
		1	56	6	6	.86	14	6	4	.76	70	6	6	.87
Test 9-Metaphor Continuation	R	1	56	6	4	.58	15	5	5[c]	-.10	71	5	5	.64
		2	56	6	4	.42	16	6	5	-.10	72	6	6[b]	.56
	P	2	56	6	4[b]	.60	15	6	4	.68	71	6	4[b]	.78
		1	56	6	5	.61	15	6	5	.59	71	6	5[b]	.62

MC Test Battery[e]	R	1	54	84	50	.77	13	80	59	.71	66	77	72	.93
		2	55	82	54	.78	12	82	58	.64	67	78	71	.94
	P	1	56	43	33	.82	13	44	35	.56	69	43	37	.91
		2	56	41	32	.85	14	42	36	.37	70	40	37	.93
	R&P	1	54	127	83	.86	11	124	94	.68	64	120	109	.95
		2	55	123	86	.89	11	124	94	.30	66	118	108	.96
WAT	R	1&2	111	40	40	.85	30	40	40	.88	141	40	40	.92

Note. Key to column headings: R/P = receptive or productive; N = number of participants; K1 = number of test items at start of the instrument reliability analysis; K2 = number of test items at end of the instrument reliability analysis; α = Cronbach's alpha.

[a] Since the item analysis resulted in the preliminary deletion of rogue items in each data file separately, there are some differences in 'start item' values between the three files.

[b] Items chosen to ensure no differences between G1 & G2 scores.

[c] Items for which all participants scored full marks retained.

[d] Reasons [b] and [c].

[e] Reliability estimates of all items retained above from tests 1-9.

Table 5.6 *Instrument Reliability of MC Test Battery: Tests-within-Battery*

Data file	Mode (R/P)	Participants (*N*)[a]	MC tests (*K*)[b]	Cronbach's alpha (α)
NNS	Receptive	109	9	.58
	Productive	112	6	.70
	Receptive & productive	109	15	.77
NS	Receptive	25	9	.28
	Productive	27	6	.36
	Receptive & productive	22	15	.13
NNS+NS	Receptive	133	9	.87
	Productive	139	6	.87
	Receptive & productive	130	15	.92

[a] Listwise deletion.

[b] All MC tests (1-9) except Test 4-Topic/Vehicle-P.

5.2.6.1 Results

Tables 5.5 and 5.6 show the final numbers of items retained for each group's test after the instrument reliability analysis had been conducted on each data file. The data reveal a wide range of internal consistency estimates between tests in the MC Test Battery. Alphas were highest in the NNS+NS data file, and for the MC Test Battery (overall) and WAT (all files), and lowest (or indeterminate) for several MC tests (all files). The 30 alphas of MC tests (groups 1 and 2, receptive and productive) varied dramatically between the NNS data file (*M* = .49, *SD* = .18, *Mdn* = .50, *IQR* = .21), NS data file (*M* = .30, *SD* = .35, *Mdn* = .39, *IQR* = .65) and NNS+NS data file (*M* = .63, *SD* = .15, *Mdn* = .63, *IQR* = .22). Despite the variation, one can observe that alphas are general highest for Test 3-Vehicle Acceptability-R, Test 8-Idiom Extension (both –R and –P), and Test 2-Metaphor Layering-R.

Two methods were used to estimate the reliability of MC Test Battery (overall): Method 1, Items-within-battery - calculating alpha coefficients as the internal consistency of all 50+ receptive, 32+ productive, and 83+ receptive and productive item scores[52] comprising the MC Test Battery (MC Test Battery data in Table 5.5); Method 2, tests-within-battery - by calculating the internal consistency of all 9 receptive, 6 productive, and 15 receptive and productive test scores comprising the MC Test Battery (see Table 5.6).[53]

Method 1 leads to high reliability MC Test Battery alphas (.77 or above) in the NNS and NNS+NS data, while Method 2 estimates are high but on the whole slightly lower. In the NS data file, estimates are significantly lower, and again lowest by method 2. Although it was possible to delete more items in the NS data file to obtain an MC Test Battery with higher alphas (.74 or above), these data were not needed for the EFA or MANOVA in subsequent chapters and so this step was not implemented.

5.2.6.2 Do any tests need to be removed due to low instrument reliability?

A decision was made that no tests in the present study should be deleted based on instrument reliability estimates. This decision may appear questionable given that, although MC Test Battery (overall) and WAT estimates are high, alpha values for items-within-tests in the three data files suggest that many of the MC tests have retained sets of items with unacceptably low internal consistency. The issue of instrument reliability in the present study is a complicated one.

[52] As shown in Appendix F, the final number of items retained differs between data files. Thus, "+" is used to indicate a minimum of 50, 32, and 83 items retained in the NNS data file, but more in other data files.
[53] The second of these methods is equivalent to the way in which Loewen, Li, Fei, Thompson, Nakatsukasa, Ahn and Chen (2009) reported the internal consistency of items in a questionnaire that was later factor analysed.

While some researchers might favour blindly cutting tests to satisfy generic rules-of-thumb, this is not the approach used here, and the complexities surrounding instrument reliability are taken up in the next chapter.

At this point in the data cleaning, the 'final' sets of items retained in the NNS, NS and NNS+NS data files have been presented. The reader can view these in Appendix F and will notice that in order to make tests as internally consistent as possible, different items were retained within each data file.

5.2.7 Interrater and intrarater reliability analyses

In order to help finalise the scoring criteria and provide information for answering the first research question, interrater and intrarater reliability analyses of scoring decisions for limited production responses were conducted. These analyses were not used to delete items, but are reported here to make it easier for the reader to compare the approaches and results of the different types of reliability analyses.

In total, three raters scored (NNS and NS) responses to 78 questions requiring test takers to write (i.e., type) a response. Table 5.7 (below) presents mean percentage agreement and weighted kappa coefficients (*Kw*) at three scoring stages for each test (1 x receptive test, 6 x productive tests), along with standard deviations, number of items and bootstrap statistics. For Test 2-Metaphor Layering-R (Aa), data are based on 143 NNS and NS responses to the same six (receptive) questions. For all other tests data are based on 71 NNS and NS responses to group 1's six productive items (i.e., items 7-12), and 72 NNS and NS responses to group 2's six productive items (i.e., items 1-6). Table 5.8 (below that), presents these data for the MC Test Battery as a whole.

Both tables show that, in all cases, weighted kappa coefficients, which take chance agreements into account, are lower than corresponding percentage agreements. Both indexes show that intrarater reliability (stage 3) was higher than interrater reliability (stages 1 and 2), and that while agreement between R1 vs R2 initial scores was higher than for R1-R2 (final scores) and R3 (initial scores) for all tests except Test 5-Topic Transition-P and Test 7-Feelings-P, overall, these two sets of interrater reliability estimates (stages 1 and 2) were much the same. Estimates range from 61-94% agreement (*Kw* .49 - .86) and are highest for Test 2-Metaphor Layering-R and Test 8-Idiom Extension-P, and lowest for Test 6-Heuristic-P, Test 7-Feelings-R and Test 5-Topic Transition-P.

Table 5.7 *MC Test-by-Test Interrater and Intrarater Reliability: Limited Production Responses*

				Percentage agreement (Po)						Weighted kappa (*Kw*)					
						Bootstrap[a]						Bootstrap[a]			
								95% CI of *M*						95% CI of *M*	
Test	Scoring stage	*N*[b]	*K*	*M*	*SD*	Bias	*SE*	Lower	Upper	*M*	*SD*	Bias	*SE*	Lower	Upper
Test 2-Metaphor Layering-R	1) Interrater: R1 vs. R2 (1st)	143	6	.85	0.05	0.00	0.02	.81	.89	.65	0.14	0.00	0.05	.56	.75
	2) Interrater: R1-R2 (final) vs R3 (1st)	143	6	.81	0.16	0.00	0.06	.68	.91	.60	0.25	0.00	0.09	.40	.76
	3) Intrarater: R1-R2 (final) vs R1 (2nd)	143	6	.94	0.04	0.00	0.01	.91	.97	.86	0.08	0.00	0.03	.80	.91
Test 4-Topic/Vehicle-P	1) Interrater: R1 vs. R2 (1st)	143	12	.75	0.12	0.00	0.03	.69	.82	.67	0.14	0.00	0.04	.60	.74
	2) Interrater: R1-R2 (final) vs R3 (1st)	143	12	.70	0.14	0.00	0.04	.63	.77	.59	0.14	0.00	0.04	.51	.66
	3) Intrarater: R1-R2 (final) vs R1 (2nd)	143	12	.86	0.07	0.00	0.02	.82	.90	.82	0.09	0.00	0.02	.78	.87
Test 5-Topic Transition-P	1) Interrater: R1 vs. R2 (1st)	143	12	.63	0.16	0.00	0.05	.54	.72	.52	0.21	0.00	0.06	.40	.64
	2) Interrater: R1-R2 (final) vs R3 (1st)	143	12	.76	0.10	0.00	0.03	.71	.81	.72	0.12	0.00	0.03	.65	.78
	3) Intrarater: R1-R2 (final) vs R1 (2nd)	143	12	.80	0.06	0.00	0.02	.76	.83	.76	0.06	0.00	0.02	.73	.80
Test 6-Heuristic-P	1) Interrater: R1 vs. R2 (1st)	143	12	.68	0.08	0.00	0.02	.63	.72	.58	0.10	0.00	0.03	.53	.64
	2) Interrater: R1-R2 (final) vs R3 (1st)	143	12	.66	0.10	0.00	0.03	.61	.71	.56	0.15	0.00	0.04	.48	.64
	3) Intrarater: R1-R2 (final) vs R1 (2nd)	143	12	.81	0.04	0.00	0.01	.79	.83	.75	0.06	0.00	0.02	.72	.78
Test 7-Feelings-P	1) Interrater: R1 vs. R2 (1st)	143	12	.61	0.15	0.00	0.04	.53	.69	.49	0.20	0.00	0.05	.40	.60
	2) Interrater: R1-R2 (final) vs R3 (1st)	143	12	.70	0.11	0.00	0.03	.64	.76	.59	0.12	0.00	0.03	.54	.66
	3) Intrarater: R1-R2 (final) vs R1 (2nd)	143	12	.84	0.06	0.00	0.02	.81	.88	.77	0.11	0.00	0.03	.71	.82
Test 8-Idiom Extension-P	1) Interrater: R1 vs. R2 (1st)	143	12	.82	0.07	0.00	0.02	.78	.85	.75	0.10	0.00	0.03	.69	.80
	2) Interrater: R1-R2 (final) vs R3 (1st)	143	12	.76	0.08	0.00	0.02	.71	.80	.71	0.11	0.00	0.03	.64	.77
	3) Intrarater: R1-R2 (final) vs R1 (2nd)	143	12	.88	0.05	0.00	0.01	.85	.91	.86	0.06	0.00	0.02	.82	.89
Test 9-Metaphor Layering-P	1) Interrater: R1 vs. R2 (1st)	143	12	.73	0.12	0.00	0.03	.66	.80	.64	0.16	0.00	0.05	.55	.72
	2) Interrater: R1-R2 (final) vs R3 (1st)	143	12	.69	0.12	0.00	0.03	.63	.76	.61	0.13	0.00	0.04	.54	.68
	3) Intrarater: R1-R2 (final) vs R1 (2nd)	143	12	.85	0.06	0.00	0.02	.82	.89	.83	0.07	0.00	0.02	.79	.86

[a] Bootstrap results are based on 1,000 bootstrap samples.

[b] Although groups 1 and 2 encountered different items for tests 4, 5, 6, 7, 8, and 9, Independent samples *t*-tests showed no statistically significant differences (at the .01 level) between estimates for group 1 and 2 items for these tests, suggesting it was viable to compute a mean from all 12 coefficients.

Table 5.8 *MC Test Battery Mean Interrater and Intrarater Reliability Estimates: Limited Production Responses*

						Bootstrap[a]			
								95% Confidence interval of mean	
Scoring stage	*N*	*K*	*Po*	*kw*	*SD*	Bias	*SE*	Lower	Upper
1) Interrater: R1 vs. R2 (1st)	143	78	.71	_	0.14	0.00	0.02	.68	.75
	143	78	_	.61	0.17	0.00	0.02	.57	.65
2) Interrater: R1-R2 (final) vs R3 (1st)	143	78	.72	_	0.12	0.00	0.01	.69	.75
	143	78	_	.63	0.15	0.00	0.02	.60	.66
3) Intrarater: R1-R2 (final) vs R1 (2nd)	143	78	.85	_	0.07	0.00	0.01	.83	.86
	143	78	_	.8	0.08	0.00	0.01	.78	.82

[a] Bootstrap results are based on 1,000 bootstrap samples.

Data for the MC Test Battery as a whole show that between-rater agreement was 71-72% (*Kw* = .61 to .63) at stages 1 and 2, and within- rater agreement was 85% (*Kw* = .80). These findings are discussed further in the next chapter.

5.2.8 Version parity analysis: Merging group 1 and group 2's MC Test Battery scores and converting to mean percentages

In this section, an analysis confirming the parity of group 1 scores (MC Test Battery version 1) and group 2 scores (MC Test Battery version 2) is reported. Before combing both group 1 and 2's receptive scores on one receptive variable and their productive scores on one productive variable for use in the Exploratory Factor Analysis, MANOVA (Chapter 7) and regression analyses (Chapter 9), it was necessary to measure and confirm parity between the two versions of each test.

The principles used to select items and create tests meant that MC Test Battery versions 1 and 2 were theoretically and qualitatively equivalent. Participants' scores for each test were calculated from the final items retained and converted to percentages. Since Kolmogorov-Smirnov and Shapiro-Wilk tests showed that at least one of the two groups' percentage scores were nonnormally distributed for each test, Mann-Whitney U test was used to compare means. A total of 42 comparison were made comprising 14 tests (receptive and productive tests 1, 4, 5, 6, 7, 8, and 9) in three data files (NNS, NS, and NNS+NS). The results revealed no statistically significant differences (at the .05 level) in 33 comparisons and statistically significant differences in nine comparisons[54] (at the .05 level). In 33 out of 42 cases therefore, statistical parity between group 1 and group 2's tests had been achieved by the items retained thus far. In the remaining nine cases, Cronbach's alpha coefficients and further Mann-Whitney U tests were used to identify the most internally consistent sets of items with no statistically significant differences. Lower and upper bound 95% confidence intervals for Cohen's *d* effect sizes[55] for all 42 test version comparisons were found to pass through zero, indicating statistical equivalence. These tests are marked with [b] in the instrument reliability analysis (section 5.2.6.1).

Now that this analysis had confirmed version parity, group 1 and 2's scores were combined to form single receptive and productive variables for each test to be used for further analyses in chapters 6 and 8. In the remainder of this chapter, in order to answer the second research question, descriptive statistics of the merged test scores are presented.

---

[54] These were, in the NNS data file: Test 1-Phrasal Verbs-P, Test 7-Feelings-R, Test 9-Metaphor Continuation-P; In the NS data file: Test 6-Heuristic-R; In the NS+NNS data file: Test 1-Phrasal Verbs-R, Test 1-Phrasal Verbs-P, Test 7-Feelings-R, Test 9-Metaphor Continuation-R, Test 9-Metaphor Continuation-P.

[55] Cohen's *d* calculated using Becker's (2000) Effect Size Calculators (http://www.uccs.edu/~lbecker/), Confidence Intervals calculated using syntax developed by Jeromy Anglim (2016) for R package 'compute.es' (Del Re, 2013).

5.3 Descriptive statistics

5.3.1 Results

Descriptive statistics were generated in order to be able to answer research question 2, which asked about apparent differences between NNS and NS MC test scores. Table 5.9 lists the NNS, NS, and (since NNS and NS combined data are used in later analyses) NNS+NS group mean and median scores for each MC test, the MC Test Battery, vocabulary, and overall and component L2 proficiency tests (NNSs only). The number of participants (*N*), distributions, and the high-to-low rank score orders for each type of test (i.e., MC, vocabulary, or proficiency) are also shown. Percentages scores for Test 2-Metaphor Layering-R were calculated giving equal weighting to 'Aa', 'Ab', 'B' and 'C'-type questions (see section 4.3.5.3). Parts A and B of Test 3-Vehicle Acceptability-R (see section 4.3.5.4) were also weighted equally. MC Test Battery-R and –P scores were calculated as participants' mean percentage score for all receptive and productive tests respectively; MC Test Battery-R&P scores were the mean of these two. If a participant had been deleted as an outlier for any receptive or productive MC test, their data were not included in the overall MC Test Battery scores. Maximum scores for the other tests were 10,000 words (VYesNo), 160 associates (WAT), 120 points including 20 per CEFR level (OOPT), and 9.0 (IELTS). Most scores were nonnormally distributed, and so median and interquartile ranges are the most appropriate measures of average and spread. For normally distributed scores (indicated as exceptions), mean and standard deviation are the most appropriate measures of average and spread. Table 5.10 presents the MC Test Battery-P and –R&P scores, recalculated after the removal of Test 4-Topic/Vehicle-P from the data (see section 5.3.2 below). Figures 5.4 and 5.5 show box-and-whisker plots of the NNS and NS MC test and MC Test Battery scores.[56]

For MC tests and the MC Test Battery the results show that, as expected, the NS group scored higher than the NNS group for all MC and vocabulary tests. NNS scores were notably higher for Test 5-Topic Transition-R than any other test, and lowest for Test 8-Idiom Extension-R and –P. Variation in NNS scores was highest for Test 8-Idiom Extension-P. The NSs, on the other hand, scored highest for Test 1-Phrasal Verbs-R and –P, Test 5-Topic Transition-R and Test 7-Feelings-R. NS scores were lowest and most varied for Test 8-Idiom Extension-R and Test 4-Topic/Vehicle-P. The NNSs attained higher receptive than productive scores for the overall battery variables and all tests except Test 1-Phrasal Verbs-R and –P. The NSs also scored higher in the receptive mode, but had Test 8-Idiom Extension-P and Test 9-Metaphor Continuation-P as exceptions to this pattern.

[56] Mean = crosses within boxes; Median = horizontal lines within or on (short) edge of boxes; IQR = boxes; Q1 and Q3 = bottom and top (short) edges of boxes; outliers (> Q3 + 1.5 times IQR or < Q1 – 1.5 times IQR) = points.

Table 5.9 *Descriptive Statistics of All Tests*

	NNS data file							NS data file							NNS+NS data file						
Test/Variable	*N*	*M*(%)	*Mdn*(%)	*M* rank	*Mdn* rank	*SD*	*IQR*	*N*	*M*(%)	*Mdn*(%)	*M rank*	*Mdn rank*	*SD*	*IQR*	*N*	*M*(%)	*Mdn*(%)	*M* rank	*Mdn* rank	*SD*	*IQR*
T1-Phrasal Verbs-R	112	58.4	58.3	6	6	24.8	25	30	99.3	100	1	1	2.5	0	142	59.0	55.6	8	9	26.3	37.5
T1-Phrasal Verbs-P	112	64.3	75	3	2	25.5	25	30	96.7	100	3	1	6.6	2.5	142	54.4	50	11	10	28.0	39.8
T2-Metaphor Layering-R[a]	111	57.7	55	7	7	23.9	43.8	29	89.9	100	7	1	21.5	11.1	140	61.7	57.9	5	8	20.9	29.1
T3-Vehicle Acceptability-R	112	50.2	54.7	1	8	26.0	38.1	30	91.6	100	6	1	13.0	16.7	142	49.3	43.8	12	14	27.7	34.4
T4-Topic/Vehicle-R	111	60.6	50	4	9	26.7	25	31	83.9	100	9	1	22.9	25	142	65.7	75	4	1	27.6	50
T4-Topic/Vehicle-P	112	53.1	50	9	9	24.1	37.5	31	70.6	75	15	15	29.0	50	143	56.9	62.5	10	5	26.1	25
T5-Topic Transition-R	111	70.5	75	1	2	26.2	50	30	98.3	100	2	1	5.1	0	140	76.8	75	1	1	25.2	50
T5-Topic Transition-P	112	38.3	37.5	14	14	25.8	41.7	30	81.7	87.5	13	12	17.9	15.6	142	47.8	50	14	10	28.5	50
T6-Heuristic-R	112	66.6	80	2	1	25.4	40	30	94.4	100	5	1	10.1	16.7	142	70.1	75	2	1	27.6	50
T6-Heuristic-P	112	55.5	62.5	8	4	25.8	37.5	31	82.3	87.5	12	12	18.2	25	143	59.5	60	7	6	26.1	40
T7-Feelings-R	112	59.8	60	5	5	23.5	40	30	96.1	100	4	1	8.4	0	142	66.0	66.7	3	4	25.5	33.3
T7-Feelings-P	112	47.4	43.8	12	13	25.3	37.5	31	83.5	100	11	1	20.3	25	143	58.7	50	9	10	27.0	42.5
T8-Idiom Extension-R	112	24.3	25	16	15	29.7	43.8	31	64.5	75	16	15	33.2	60	143	35.1	20	16	16	32.6	60
T8-Idiom Extension-P	112	31.9	16.7	15	16	32.5	58.3	30	83.6	89.6	10	10	22.3	25	142	42.8	41.7	15	15	36.7	75
T9-Metaphor Continuation-R	112	49.3	50	11	9	29.9	50	31	86.6	83.3	8	14	11.8	20	143	59.6	60	6	6	28.0	50
T9-Metaphor Continuation-P	112	41.7	45	13	12	27.9	50	30	79.7	88.8	14	11	25.3	30	142	48.9	50	13	10	31.6	50
MC Test Battery-R[b]	109	55.6	55.7	1	1	12.6	17.3	25	90.3	91.7	1	1	6.3	9.2	133	59.8	55.5	1	1	18.6	23.1
MC Test Battery-P[b]	112	47.5	48.8	3	3	16.3	23.1	27	85.0	87.9	3	3	8.0	12.5	139	52.4	48.0	3	3	22.0	32.0
MC Test Battery-R&P[b]	109	51.8	51.4	2	2	12.9	17.6	22	88.4	90.1	2	2	4.9	9.4	130	55.7	51.0	2	2	18.9	22.0
VYesNo[cd]	111	5918	6029	2	2	1187	1632	30	8902	9384	2	2	1242	1240	141	6553	6421	2	2	1711	2122
Word Associates Test[ed]	111	126	126	1	1	11	15	30	148	152	1	1	8	10	141	131	130	1	1	14	19
OOPT (overall)[c]	112	66.98	68	2	2	13.91	23.5														
OOPT Use of English[a]	112	67.57	69	1	1	15.02	19														
OOPT Listening	112	66.30	67	3	3	17.44	27.75														
IELTS (overall)	111	6.64	6.5	3	3	0.51	0.5														
IELTS Reading	111	7.07	7	1	1	0.87	1														
IELTS Writing	111	6.05	6	5	4	0.47	0														
IELTS Speaking	111	6.19	6	4	4	0.55	0.5														
IELTS Listening	111	7.04	7	2	1	0.94	1														

[a] NNS data normally distributed, Kolmogorov-Smirnov test Sig. > .05.

[b] NNS and NS data normally distributed, Kolmogorov-Smirnov and Shapiro-Wilk tests Sig. > .05.

[c] NNS data normally distributed, Kolmogorov-Smirnov and Shapiro-Wilk test Sig. > .05.

[d] VYesNo (scores out of 10,000) and WAT (scores out of 160) converted to percentages for ranking of means and medians.

[e] NNS and NNS+NS data normally distributed, Kolmogorov-Smirnov Sig. > .05 in both files, Shapiro-Wilk Sig. > .05 in NNS data file and =.050 in NNS+NS data file.

Table 5.10 *Metaphoric Competence Variables with Test 4-Topic/Vehicle-P Removed*

	NNS data file							NS data file							NNS+NS data file						
Test/Variable	*N*	*M(%)*	*Mdn(%)*	*M rank*	*Mdn rank*	*SD*	*IQR*	*N*	*M(%)*	*Mdn(%)*	*M rank*	*Mdn rank*	*SD*	*IQR*	*N*	*M(%)*	*Mdn(%)*	*M rank*	*Mdn rank*	*SD*	*IQR*
MC Test Battery-P[a]	112	46.5	47.9	2	2	17.2	24.2	27	86.3	87.6	2	2	8.7	13.3	139	51.5	48.7	2	2	23.4	32.5
MC Test Battery-R&P[a]	109	51.4	50.5	1	1	13.2	19.4	22	89.3	89.6	1	1	4.9	7.4	130	55.2	51.2	1	1	19.6	22.1

[a] NNS and NS data normally distributed, Kolmogorov-Smirnov and Shapiro-Wilk tests Sig. > .05.

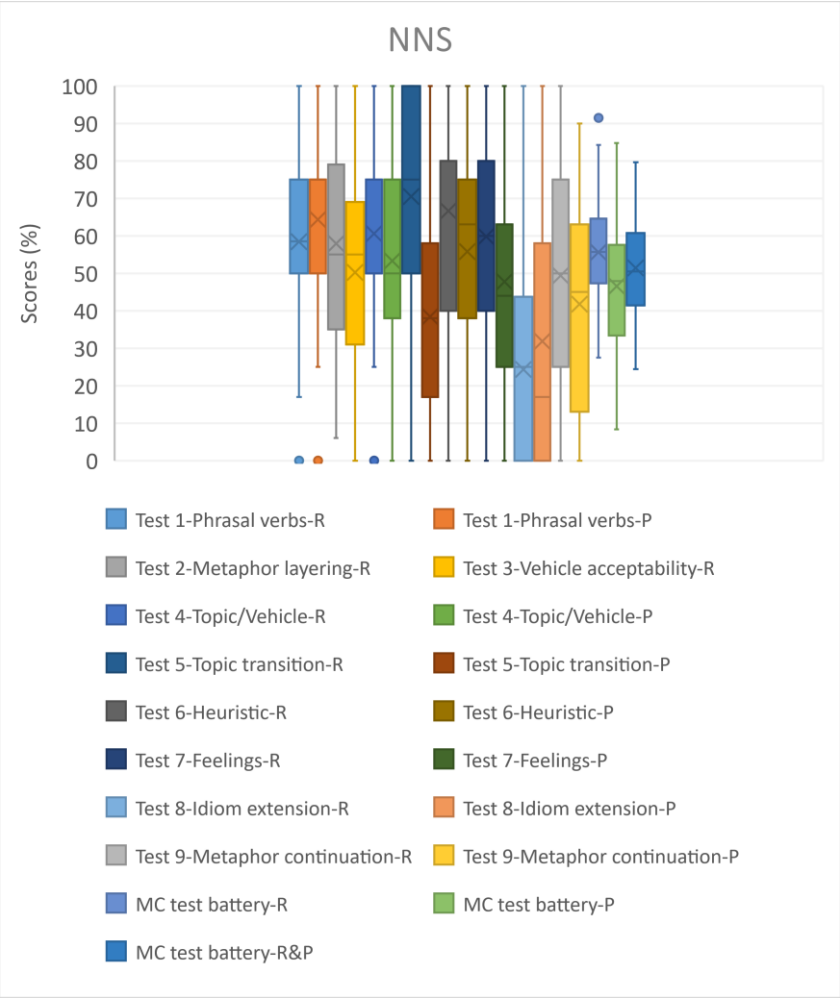


Figure 5.3 NNS descriptive statistics MC Test Battery

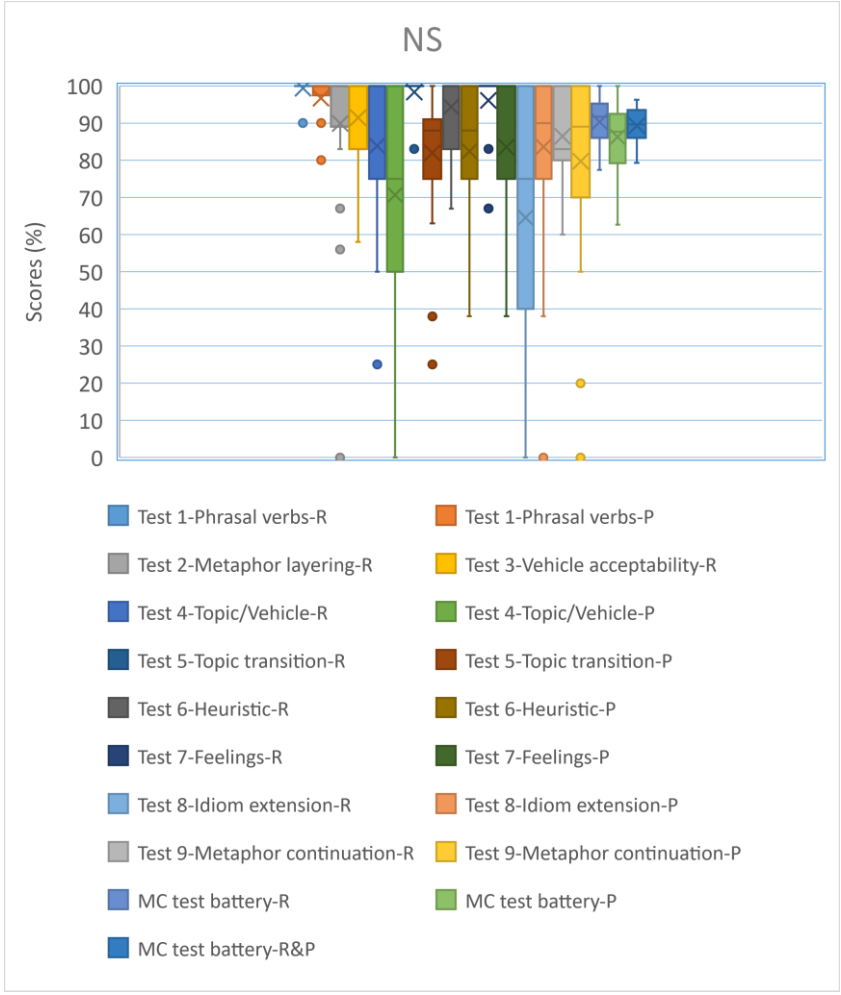


Figure 5.4 NS descriptive statistics MC Test Battery

For vocabulary tests, VYesNo scores show that the NNSs demonstrated an average vocabulary size of around 6,000 words whereas the NSs scored around 9,000 words. The WAT scores showed that, on average, the NNSs had a depth of lexical network covering close to 80% of word associates, whereas the NSs recognised over 90% of associates. Notably, both the NNSs and NSs scored closer to full marks for the WAT than VYesNo.

For L2 proficiency, the NNSs scored slightly higher for the Use of English than Listening section of the OOPT, averaging CEFR B2 level. The reported IELTS scores suggested higher scores for reading and listening than other strands. Using the British Council's Common European Framework Equivalencies (2017), IELTS scores ranging 4.0 – 5.0 were indexed as CEFR B1, 5.5 – 6.5 as B2, 7.0 – 8.0 as C1, and 8.5 – 9 as C2. These were compared with the OOPT test scores, which the administration interface had automatically indexed as CEFR levels[57] (see also Pollitt, 2016, p. 9). A direct comparison of proficiency tests is complicated by the fact that Mean, Median categorise the NNS group's IELTS (overall) and IELTS Reading scores as B2, whereas the Mode places these at C1. However, if one works with just the Mean and Median, the key finding is that although both the overall OOPT and reported IELTS scores show that the NNSs had an 'average' CEFR level of B2, Mean group listening scores are different for both tests, namely C1 for IELTS and B2 for the listening section of the OOPT. As shall be seen in chapters 6 and 8, the differences between these two tests become even more significant with regard to L2 metaphoric competence.

5.3.2 Do any MC tests need to be removed due to low NS group scores?

While the NSs attained high group scores for most tests (*M* and *Mdn* > 80%), two tests do not fit this pattern: Test 4-Topic/Vehicle-P (*M* = 70.56, *SD* = 28.97; *Mdn* = 75, IQR = 50), and Test 8-Idiom Extension-R (*M* = 64.52, *SD* = 33.18; *Mdn* = 75, IQR = 60).

Test 4-Topic/Vehicle-P required participants to supply words to complete analogies such as 'CCTV cameras are the _____ of the building'. To complete these tasks, test takers essentially had to think how 'CCTV cameras' might relate to 'buildings', and come up with something from another area of life (e.g., a concrete noun) comparable to 'CCTV cameras' that captures some of the essence of this relationship. It is difficult to identify any obvious reason why the NSs would struggle to think up appropriate response to these examples. Upon inspection, part of the problem seems to have been numerous NS productions forming literal descriptions using abstract or nonspecific nouns such as 'security' (produced by N5B and N10B for the question above) rather than analogies employing concrete nouns. There were also several instances of nonsensical analogies such as N1B's production 'The bee hive is the brave

[57] OOPT: 0 = < CEFR A1; 1 – 20 = A1; 20 – 40 = A2; 40 – 60 = B1; 60 – 80 = B2; 80 – 100 = C1; > 100 = C2.

new world of the animal kingdom', and subjective, stretched (and sometimes impertinent) attempts at humour.[58] For these reasons, Test 4-Topic/Vehicle-P was deleted from the data and MC Test Battery-P and –R&P scores recalculated without it (see Table 5.10). All further analyses use these recalculated scores.

Test 8-Idiom Extension-R required participants to select the researcher designed *best* answer from four options for extending the literal sense of idioms. The distractor analysis of the NS responses revealed that for all items, the majority of NSs selected the researcher-designed *best* answer, and that the strongest distractor in most cases was the one that extended the everyday, figurative sense of the idiom. Unlike Test 4-Topic/Vehicle-P, the internal consistency of items for this test was comparatively high for both groups 1 and 2. Responses to the productive test of idiom extension revealed that even those NSs who struggled to recognise *best* answers in the receptive test were able to produce successful idiom extensions of literal senses, suggesting that low NS receptive scores are not likely to be attributable to test takers simply misunderstanding the task.

Consequently, Test 8-Idiom Extension-R was retained in the MC Test Battery for further analyses.

5.4 Chapter summary

The focus of this chapter was on investigating:

a) the extent to which (L1 and L2) metaphoric competence can be reliably elicited and measured (RQ1)
b) how MC test scores appear to differ for English NNSs (L1 Chinese) and NSs (RQ2)

Data cleaning revealed 'rogue' items and items which should be deleted to improve the internal consistency of tests. In many cases, the items retained and deleted were different in the NNS and NS data files. **Rating scale outliers** signifying high variation in NS acceptability judgements were identified in Test 3-Vehicle Acceptability-R and Test 4-Topic/Vehicle-R, leading to the removal of 10 (out of 28) and 1 (out of 12) items respectively from all data files. **Participant outliers** implying possible question-skipping were identified in each data file leading to the removal of 19 (6 NNSs, 13 NSs) test scores. The **item analysis** resulted in the deletion of 19 items from the data that were poor discriminators, too difficult (NS data file only) or for which the NNSs scored higher than the NSs. The results of the **distractor analysis** showed that distractors, on the whole, were better at luring low ability NNSs than low ability NSs. Distractors for both groups performed comparatively well for the WAT and Test 9-Metaphor Continuation-R, and

[58] One (repeatable) example of this is N6A's response 'Chemical elements are the Bob Holness of life'. Although raters were familiar with Bob Holness (a presenter of the British TV gameshow 'Blockbusters'), they decided that the response did not make logical sense, and so scored it '0' (incorrect).

poorly for Test 4-Topic/Vehicle-R.

The **instrument reliability analysis** showed a wide range of internal consistency estimates for MC tests and the WAT with higher alphas for tests in the NNS+NS data file, the MC Test Battery (overall) and WAT (all files), and lowest (or indeterminate) for several MC tests (all files). Despite variation across the three data files, items were most internally consistent for Test 3-Vehicle Acceptability-R, Test 8-Idiom Extension (both –R and –P), and Test 2-Metaphor Layering-R. Further reliability analyses showed that **Intrarater reliability** estimates were higher than two sets of **interrater reliability** estimates involving three raters. Generally, these estimates were highest for Test 2-Metaphor Layering-R and Test 8-Idiom Extension-P, and lowest for Test 6-Heuristic-P, Test 7-Feelings-R and Test 5-Topic Transition-P. Since they take change agreements into account, weighted Kappa coefficients were lower than percentage agreements in all cases. The **version parity analysis** confirmed the statistical equivalence of group 1 and group 2's scores, which for each test, were combined to form one larger dataset.

Finally, the **descriptive statistics** showed that the receptive rather than productive mode generally yielded higher test scores for both NNS and NS groups. NNS group scores were highest for 5-Topic Transition-R and lowest for Test 8-Idiom Extension-R and –P, whereas the NSs attained their highest scores for Test 1-Phrasal Verbs-R and –P, Test 5-Topic Transition-R and Test 7-Feelings-R, and their lowest scores for Test 8-Idiom Extension-R and Test 4-Topic/Vehicle-P. Due to low NS scores and productions which indicate the skill was not being engaged in by many test takers, Test 4-Topic/Vehicle-P was deleted from the MC Test Battery and overall –R, -P and –R&P scores for the battery recalculated.

Chapter 6: Discussion of Analysis 1

6.1 Introduction

This chapter will discuss the key findings of **Analysis 1: Development and reliability of the MC Test Battery, descriptive statistics**, in relation to previous metaphoric competence and other research. The discussion is structured into two main parts corresponding to the first two research questions, and subsections based on important themes emerging from the analyses in the previous chapter.

6.2 RQ1: To what extent can (L1 and L2) metaphoric competence be reliably elicited and measured?

To answer the first research question, metaphoric competence was conceptualised as the metaphor-related skills described by Low (1988) and Littlemore and Low (Littlemore & Low, 2006a). A large battery of MC tests was then developed, piloted, and administered to 112 L1 Chinese NNSs of English and 31 English NSs along with tests of vocabulary knowledge and (for the NNSs only), L2 proficiency. Finally, the data were 'cleaned' using analyses to identify rating scale, participant and item outliers and instrument reliability and version parity analyses. Further aspects of reliability were assessed via a distractor analysis and interrater and intrarater reliability analyses, although these analyses were not used for deleting items. The discussion of the first research question below is conducted under four general themes: statistical reliability, operational challenges, NS variation and the use of the multiple-choice format.

6.2.1 Statistical reliability of the MC Test Battery

The first consideration in assessing the extent to which (L1 and L2) metaphoric competence can be reliably measured concerns the heart of the data cleaning analyses, statistical reliability of the MC Test Battery. For convenience of the discussion in this chapter (and subsequent ones), I use the following adjectives to refer to the Cronbach's alpha values observed: $\alpha > .8$ ('very high'); $.8 > \alpha \geq .7$ ('high'); $.7 > \alpha \geq .5$ 'moderate'); $\alpha < .5$ 'low'). Importantly, these are ad-hoc terms and *not* externally stipulated benchmarks, used only for categorising observed alpha values in relation to one another. They do not, therefore, indicate my endorsement of a 'rule-of-thumb' approach to reliability, which has been shown to be a far more complex issue. Percentage agreement and weighted kappa coefficients (inter- and intrarater reliability) are referred to in relative terms only (e.g., higher or lower than X).

In the previous chapter, it was shown that as one large L2 metaphoric competence instrument, the MC Test Battery is a highly reliable measure, evidenced by 'high' to 'very high'

Cronbach's alpha estimates for the NNSs for items-within-battery (α = .77 to .89) and 'moderate' to 'high' estimates for tests-within-battery (α = .58 to .77). This finding is generally in line with the metaphoric competence field median of .76, although this was a reflection of both L1 and L2 MC instruments and a variety of coefficients. On the other hand, estimates for individual MC tests were generally 'low' to 'moderate', and showed quite a lot of variation (*M* = .49, *SD* = .18, *Mdn* = .50, *IQR* = .21). The extremities observed in the present study are in line with upper and lower values of .31 and .90 (Littlemore, 2001) observed in past L2 metaphoric competence research. The fact that stage 1 interrater reliability estimates (*Po* = .71, *kw* = .61) and stage 2 interrater reliability estimates (*Po* = .72, *kw* = .63) for the MC Test Battery were higher, on average, than instrument reliability, and that estimates for intrarater reliability (stage 3) (*Po* = .85, *kw* = .80) were higher than instrument and interrater reliability aligns with the SLA field more generally (Plonsky & Derrick, 2016).

It should also be noted here the WAT appeared to be slightly more internally consistent for the NSs (α = .88) than the NNSs (α = .84). Given Read's (1993) point that native speakers have more stable patterns of word association reflecting richer lexical and semantic networks, this would probably be expected. The 'high' WAT alpha values for the NNSs in the present study does, however, undermine the assertion that NNSs have much less stable networks of word association. Importantly, the 'high' values observed may also be a product of the mode in which depth of lexical network was tested. Although it is an empirical question, one would expect a productive measure of word association, particularly one consisting of much lower frequency words, to yield lower internal consistency estimates for both NNSs and NSs (Fitzpatrick, 2007).

As noted in Chapter 3, internal consistency estimates can be expected to be lower for tests with fewer than 10 items; tests that measure psychological constructs, various constructs (A Field, 2013; P. Kline, 1999; Pallant, 2013), or broad rather than narrow constructs (Peters, 2014); samples of participants with lower L2 proficiency; and lower for certain SLA subdomains (Plonsky & Derrick, 2016). One other reason why researcher-developed tests (such as these) may be less reliable than standardised tests is that reliability for the latter is sometimes artificially inflated via the inclusion of more high ability test takers to increase inter-item correlations (Purpura, 2009). In the case of the present study, 'low' reliability of some MC tests was most likely due to a combination of several factors. Probably the most substantial of these is the small number of items for each test, which would explain why the instrument reliability is substantially higher for the MC Test Battery scores, which involve (on average) more than 11 times as many items. A second influential factor was probably that tests did not measure singular constructs. This issue is explored further in the following two chapters.

Test 8-Idiom Extension-R and -P had by far the most internally consistent set of items, and for productive responses, higher interrater and intrarater reliability estimates than for other

productive MC tests. 'High' instrument reliability for this test is probably explained by the fact that for creative metaphor tasks such as this, scores are frequently skewed towards the lower end of scales (Beaty & Silvia, 2013). In other words, if a learner possesses a creative ability to extend the literal sense of idioms, they tend to be able to exercise it consistently for all items. Conversely, a test taker who struggles to extend one idiom, is likely to struggle to extend more. By comparison, interrater and intrarater estimates are likely to have been higher than for other MC tests because the scoring categories for this test were much clearer than for others. The stipulation that participants needed to extend the literal sense of idioms to gain marks was, it seems, more consistently implemented than criteria for any other test.

Another possible explanation for the 'high' internal consistency for Test 8-Idiom Extension-R and –P relates to task approach rather than competence in this area per se. Specifically, some participants may have simply preferred to extend the figurative sense of idioms (thus earning marks consistently) and others not. Strictly speaking, the instructions for this task did not state *you need to extend the idiom by referring to the literal sense.* Rather, in order to avoid test takers gaining marks by simply spotting or producing any literal sense, test takers were left to infer from the example and explanation provided that an extension of the idiom's literal sense was required and rewarded for recognising this. Whether or not the test would remain reliable with more deductive instructions is an empirical question. Because a close examination of the data revealed that only eight out of 112 NNSs (3 from group 1 and 8 from group 2) did not select or produce a single response or distractor that extended the literal sense of idioms, it is unlikely that task approach had a large influence on the consistency of scores.

In addition to its 'high' reliability, the major finding for Test 8-Idiom Extension-R and -P is that contrary to Littlemore and Low's (2006a) advanced learners, participants in the present study did very well to engage in this skill at all. This is especially interesting given that most could not be considered 'advanced' learners of English. This discrepancy is probably due to differences between the authors' task and those in Test 8-Idiom Extension-R and –P. Littlemore and Low's (2006a) approach of showing Prodromou's (2003) manipulated idioms to participants and then asking them to come up with their own adaptions may have left learners feeling somewhat thrown in the deep end. In the present study, all MC tests were preceded by instructions using as little technical language as possible, examples and explanations. Receptive tests, generally easier than productive ones (Laufer & Goldstein, 2004), were administered first and questions order from easiest to most difficulty based on pilot study data. This is likely to have had a substantial facilitating affect.

For Test 8-Idiom Extension-R and –P, test takers were given a short explanation of what an idiom is, told that people often play with or extend idioms to emphasise something or make a joke, and given examples of an idiom in both its original and extended forms (see Chapter 4).

To add to the general point on facilitation above, it is likely that first engaging in this skill via multiple-choice questions aided the later and more difficult task (for NNSs) of creating one's own extensions. Although the multiple-choice format comes under scrutiny for several reasons (see section 6.2.4), it proved advantageous for preparing participants to produce their own literal extensions of idioms.

6.2.2 Operational challenges

A second consideration in assessing the extent to which (L1 and L2) metaphoric competence can be reliably measured concerns a series of operational challenges. These can be further subdivided into test development administration and refinement issues

6.2.2.1 Test development

In comparison with the present study, metaphoric competence research has tended to favour using a combination of newly developed and pre-existing instrumentation (Azuma, 2005; Hashemian & Nezhad, 2007; Johnson & Rosano, 1993; Littlemore, 2001; H. R. Pollio & Smith, 1980). In the present study, the combined approach allowed for investigation into how well the pre-existing vocabulary knowledge and proficiency measures could predict scores for the new MC instrument (RQ5, see Chapter 9). For all its advantages, the development of a new battery of tests brought challenges. Whether it is tacitly or explicitly acknowledged, time and funding restraints mean that researchers need to be selective in which aspects of metaphoric competence they can operationalise and measure. In the methodology chapter of the present study, details were provided about which of Low (1988) and Littlemore and Low's (2006a) skills/(sub)competences were not considered for testing, or considered and attempted but then abandoned. These decisions revealed that many of the authors' constructs, particularly those that involve large amounts of reading or writing, socially sensitive metaphors, cultural or world knowledge and non-linguistic strategies are ill-suited to elicited methods, when the aim is to develop a battery of tests, and time and the need to ensure a good amount of content coverage are important. Although logistically warranted, decisions taken to use elicited methods, administer tests in the written mode, and control the L1 of NNSs of English come with the caveats that findings may not necessarily generalise to L2 metaphoric competence in naturalistic data, the spoken mode, or different L1 groups.

6.2.2.2 Test administration

The second main operational challenge in metaphoric competence research is how to obtain data without causing participants anxiety. This is particularly relevant to L2 learners, and has been reported and addressed in a number of metaphoric competence studies (Azuma, 2005;

Johnson & Rosano, 1993; Littlemore, 2001; Zhao et al., 2014). Although it is possible to find experimental metaphoric competence research in which test taker anxiety is not reported (Aleshtar & Dowlatabadi, 2014; Hashemian & Nezhad, 2007), these appear to be exceptions. In the present study, several efforts were taken to minimise test taker anxiety (see section 4.6.4). Notwithstanding these precautions, it seemed clear to the researcher that while some NNS test takers relished the chance to participate, others felt under pressure, intimidated or fatigued by the tests. These experiences suggest that future research involving the MC Test Battery should seek to administer tests in a less time consuming way, for instance using the 'cleaned' (i.e., reduced) list of items.

6.2.2.3 Test refinement

The third main operational challenge in metaphoric competence research concerned the identification and if necessary revision, adaption or deletion of problematic items, participants or tests. In most cases, the development of new MC tests has involved substantial revisions to tests and items (Azuma, 2005; Johnson & Rosano, 1993; Littlemore, 2001; H. R. Pollio & Smith, 1980). In some studies (e.g., Hashemian & Nezhad, 2007), however, if refinements were made to new instruments, this has gone unreported. Identifying and cutting 'rogue' items, participants and tests formed a methodological and substantive part of the present study. Both versions of the MC Test Battery administered to participants contained 98 receptive and 46 productive questions. By the time data cleaning had arrived at the instrument reliability analysis, in the NNS data file MC Test Battery version 1 had been reduced to 84 receptive and 43 productive items, and version 2 to 82 receptive and 41 productive items. By the end of data cleaning (i.e., after the instrument reliability analysis, version parity analyses, and deletion of Test 4-Topic/Vehicle-P) the final set of NNS data used in the analyses in Chapters 7 and 9 comprised 50 receptive and 29 productive version 1 items and 54 receptive and 28 productive version 2 items. Put simply, for analyses involving NNSs, data cleaning resulted in the use of only 55% of the MC Test Battery version 1 items administered to participants (a 45% reduction), and 57% of MC Test Battery version 2 items (a 43% reduction). These findings are most in in line with Littlemore (2001) who reduced her tests by 29%, 33%, and 50% after piloting, and suggest that where MC is concerned, even systematically developed instruments require substantial refinement before they can give meaningful results.

The approach to data cleaning itself presented considerable challenges. What should one do with highly consistent items that are mid-range difficulty but poor discriminators? What if the most internally consistent set of items makes two versions statistically different? The eventual order and priority of analyses chosen came only after considerable thought on the implications of potential decisions. In the end, the order of rating scale outlier analysis,

participant outlier analysis, item analysis, instrument reliability analysis, version parity analysis and low overall NS test scores resulted in an iterative, rather than linear approach. This can be seen by the fact that the version parity analysis led to revisions in the instrument reliability and analysis, which needed to be checked again for statistical equivalence.

6.2.3 Variation in NS responses: A problem?

A third consideration in assessing the extent to which (L1 and L2) metaphoric competence can be reliably measured concerns variation in NS responses. The fact that the NSs varied substantially in recognising four established phrases as acceptable,[59] with standard deviations ranging from 25.68 (item 2) to 41.2 (item 21), suggest that it may not always be possible to measure NNS knowledge against NS knowledge. These findings also indicate that Low's (1988) suggestion that native speakers have a shared knowledge of the boundaries of acceptability needs to be reconsidered. Even when one can give an empirical basis for considering certain linguistic metaphors as 'commonly used', acceptance of Vehicle extension is, it seems, highly subjective.

A general reason for why Vehicles elicited mixed NS judgements may be due to the fact they were presented in decontextualised sentences. It is certainly plausible that NSs might have given more lenient Vehicle acceptability judgements if items had been encountered in an interactive discourse. Specific reasons for the lack of consensus may include the fact that '…shoot the breeze' (item 21) is more common in American than British English (see MED definition), and that '…slipped into a depression' (item 2), '…picked up a job' (item 24), and 'killer headache' (item 22) are relatively infrequent. This latter point, however, is complicated by the fact that equally low frequent forms such as 'the whole theory fell apart', which contains only one equivalent BNC-BYU entry, elicited almost complete NS consensus (*M* = 98.74, *SD* = 5.45).

At the other extreme, NSs failed to agree that six items designed to be unacceptable by the researcher[60] were in fact unacceptable, with standard deviations ranging from 26.02 (item 11) to 37.69 (item 17). For the item that elicited the most disagreement, 'he freshened his idea' (item 17), the fact that none of the 35 BNC-BYU instances of 'freshened' refer to ideas or thoughts gives little support for contemporary usage. While it was initially supposed that the different levels of agreement may be attributable to the different ages of NS participants (the hypothesis being that younger NSs would find the item more acceptable), an independent-

[59] The BNC-BYU contains 1 hit for the collocates 'slipped' and 'depression' (item 2), and for 'picked up' and 'job' (item 24). The MED confirms 'he never has time to shoot the breeze' (item 21) is a phrase, and the usage of a 'killer headache' (item 22) can be seen in numerous internet forums discussions.

[60] Item 3: 'his body went fat after a few years'; item 11: 'there was a lot of electricity between the dog and the ball'; item 14: 'he bubbled as he began shouting'; item 17: 'he freshened his ideas'; item 23: 'we solved the teased out problem very easily'; and item 27: 'I will give you a show of the ropes tomorrow'.

samples *t*-test showed no statistically significant differences (p = .616) between the age of NSs who rated the item 0-50% acceptable (n = 22, M = 40.64, SD = 17.12) and 51-100% (n = 9, M = 37.33, SD = 14.9). Further information on the ambivalence towards this item appears in the pilot think aloud comments of two NSs: "he freshened his ideas, kinda, people would definitely say it, I wouldn't say it" (rated 50%), "a bit weird again but fine" (rated 100%). These comments suggest the possible explanation that high level of variation reflects different approaches to the task, namely, some NSs rated according to what other people might find acceptable, whereas others rated according to what they themselves found acceptable.

Taken together, these findings suggest a development of Low's argument that L2 learners should be taught more on the structure and perceived rigidity of metaphor boundaries. These boundaries, it seems, vary in rigidity in opposite directions; linguistic metaphors can be both less and more acceptable than would be theoretically expected from corpus-based evidence.

6.2.4 Test format: A crucial component

A final consideration in assessing the extent to which (L1 or L2) metaphoric competence can be reliably measured concerns the issue of test format. Receptive MC tests predominantly used four-option multiple-choice questions. While this format can be found in many early (particularly L1) metaphoric competence studies (e.g., H. R. Pollio & Smith, 1980) and a few recent cases (e.g., Boers, Demecheleer, & Eyckmans, 2004), form recognition tasks such as these appear to have been less favoured for measuring receptive metaphoric competence than meaning recognition tasks (e.g., explain the meaning).

The use of the multiple-choice format brought several advantages to the present study. First, multi-choice questions are likely to have made receptive tests less intimidating to NNSs than they would have been had 'explain the meaning' tasks been used throughout. This is because form recognition is known to be easier than meaning recall (Laufer & Goldstein, 2004). Second, since selecting an option takes less time than typing a response, they allowed for the inclusion of more items and tests, meaning more of Low (1988) and Littlemore and Low's (2006a) constructs could be measured. Third, since distractors provided information about which non-target utterances low ability test takers have difficulty rejecting, using them helps fill in the picture of L2 metaphoric competence development.

Despite these advantages, the present study's use of the multiple-choice format can be criticised on several grounds. The main issue is one of generalisability; outside of the EFL classroom, when would the NNSs in the present study ever be required to recognise the *correct* metaphor from a list of several possible options? In this respect, studies that measure MC in terms of fluid, online processes (e.g., Littlemore, 2001) are arguably much more close to real

life. Second, the potential pedagogical usefulness of the MC Test Battery is limited. Littlemore and Low (2006a), for instance, were generally critical of the use of multiple-choice questions in EFL textbooks, arguing that such tasks are more suited to testing than teaching. They suggest that multiple-choice activities could be improved with appropriate input from the teacher, and (in the case of metaphorical phrasal verbs) the inclusion of diagrams showing extensions of the figurative sense of these words should be taken on board in considering how the MC Test Battery might be used in teaching. A related problem, observed in past research (e.g., Boers et al., 2014), is that exposing learners to distractors may risks encouraging them to actually acquire non-targetlike forms. On the other hand, it is only through exposure to 'negative' evidence and noticing non-targetlike forms that learners can acquire a knowledge of, to use Low's (1988) phrase, what native speakers tend not to say (Trahey & White, 1993).

Productive MC tests exclusively used the limited production format. Limited production is somewhat synonymous with sentence completion and gap-fill, and in the present study had the advantage of forcing test takers to produce specific metaphors, or engage with specific functions and structures of metaphor, thus allowing Low's (1988) and Littlemore and Low's (2006a) skills/(sub)competences to be tested. Despite this advantage, limited production is not a watertight format for eliciting metaphor. In the present study, MC tests sometimes elicited formulaic, but non-metaphorical responses. For instance, Test 5-Topic Transition item 12 yielded several NNSs and NSs productions of 'the apple never falls far from the tree', but also several instances of 'like father, like son', a formulaic, MED phrase, but not an utterance containing metaphor if MIPVU is applied. Should the formulaic response be given credit, or marked as *incorrect* because it does not involve metaphor? In this case, the issue was resolved by the fact that Littlemore and Low (2006a) had allowed for "…sayings" (p. 144) to be part of this (sub)competence, meaning non-metaphorical formulaic sequences could be deemed acceptable.

A further problem was presented by Test 9-Metaphor Continuation-P item 4, which in the receptive mode used the *best* answer '[how] is the operation unfolding?' but in the productive mode elicited numerous NNS and NS responses of 1) '[how] is it going?', and 2) '[how] is the operation going?'. The problem was that while type 2 clearly demonstrates a continuation of the metaphorical code (A JOB APPLICATION IS A SECRET AGENT OPERATION), it was uncertain whether the 'it' in type 1 responses was a pronoun referencing the operation (i.e., continuing the metaphorical code), or whether 'how's it going?' was being used as a formulaic sequence to mean something more general such as 'how is life?' or 'are you well?'. In this case, a decision to score type 2 responses as '2' (*correct*) and type 1 responses as '0' (*incorrect*), thus not giving test takers the benefit of the doubt was made. Taken together, these points illustrate that even targeted, elicited methods sometimes fail to obtain data of interest. They also highlight the fact

that productive tests pose more of a challenge than receptive tests in this respect, since there is more chance of the researcher's test 'missing the target'.

6.3 RQ2: How do metaphoric competence test scores appear to differ for a group of English NNSs (L1 Chinese) and NSs of English?

6.3.1 A basic expectation met

Research question two was phrased carefully, as an enquiry into how the NNS and NS MC test scores appear to be different between these two groups. The discussion here, therefore, does not focus on statistical differences between NNSs and NSs (covered in the next two chapters)[61] but concerns the extent to which basic expectations were met, response patterns and an attempt to account for the observed NS ceiling effects and variation.

First, what were the basic expectations? Although Low (1988) and Littlemore and Low (2006a) highlighted certain metaphor-related skills and (sub)competences that seem to be difficult for second language learners, they gave very few clues as to how these compare with one another in this respect. This made it hard to predict which of the MC tests the NNSs would find easiest and most difficult. Unfortunately, although Low (1988) had highlighted particular difficulties surrounding metaphor mixes, tuning devices and socially sensitive metaphors, none of these skills were tested in the MC Test Battery. Littlemore and Low (2006a), on the other hand, observed NNS difficulties with extending the literal sense of idioms and with metaphorical phrasal verbs, which led to the basic expectation that Test 1-Phrasal Verbs-R and –P and Test 8-Idiom Extension-R and –P would be among the most difficult in the battery for NNSs.

In different ways, both of these expectations were met. The difficulty that both NNSs and NSs had with Test 8-Idiom Extension-R and -P has been mentioned, and by observing the rank of means and medians, one can see that these were the two most difficult tests for NNSs and among the most difficult for NSs. Although the descriptive statistics appear to show that compared to other tests, the NNSs had little difficulty with Test 1-Phrasal Verbs-R (ranked 6th easiest out of 16 MC tests), and even fewer problems with Test 1-Phrasal Verbs-P (mean rank 3, median rank 2), this is misleading. This is mainly because the NSs showed clear ceiling effects for Test 1-Phrasal Verbs-R and –P (and several others), suggesting that this may be a key area of difficulty for NNSs.

6.3.2 NNS and NS differences in the rate of non-responses

Contrary to what Littlemore and Low (2006a) observed, present study participants were

[61] The enquiry in the next chapter examines NNS and NS differences in the latent (estimated) variables of metaphoric competence, rather than observed MC test scores.

surprisingly able to respond to Test 8-Idiom Extension-R and –P questions, probably due to the facilitative role of the testing format and instructions. Littlemore and Low (2006a) reported that most of their advanced learners did not write anything when asked to create their own adaptions of Prodromou's (2003) manipulated idioms. This did not receive support in the present study.

For Test 8-Idiom Extension-P, the average non-response rate (i.e., responses of "?", "no" or "no idea") was less than 9% for NNSs, with the number ranging from one non-response out of a possible 56 (item 8, '…in fact [his comment] didn't just take the cake, it ___') to 10 non-responses (item 2, 'he got such a taste of his own medicine that ___'). For the NSs, who all produced responses to items 8 and 10, only two items elicited non-responses: item 3, '…we were so stuck between a rock and a hard place that __' (two non-responses out of a possible 16), and item 7, '[let's cross that bridge]… since the decision seems likely, let's ___' (one non-response out of a possible 15). The fact that these two items also elicited relatively high numbers of NNS non-responses (5 and 7 respectively), indicates that there may be a NNS and NS connection between the specific idioms that speakers are reluctant to extend, even given ample planning time. Although this would require substantiation, it seems intuitive that if NSs struggle to think of extensions for particular idioms, this would extend to NNSs. The conclusion from the present findings, however, is limited to the specific idioms that the NNSs (and NSs) opted out of extending. The pedagogical implications of this are discussed in section 11.4.

6.3.3 Which areas of L1 metaphoric competence seem to pertain to basic and higher language cognition?

In the literature review, a connection was made between Hulstijn's (2011, 2012) theory of basic and higher language cognition (BLC and HLC), and Low's (1988) call for more research on the extent to which NSs have a shared, consistent knowledge of the acceptability of different metaphors and manipulations of metaphor. The implication was that linguistic metaphors comprehended and produced by all native speakers, may indicate areas of metaphoric competence more prototypically part of BLC, and those with substantial variation more characteristic of HLC. Due to the fact that the present study involved written rather than spoken discourse, and NSs had higher rather than mixed intellectual profiles, it was not possible to characterise MC tests as pertaining to BLC or HLC on the basis of observed NS variation. Nevertheless, if one looks at the interquartile ranges, the fact that the NSs showed ceiling effects for Test 1-Phrasal Verbs-R and –P, Test 5-Topic Transition-R and Test 7-Feelings suggests that these tests may measure more prototypically BLC areas of L1 metaphoric competence, namely, shared by all adult L1-ers, regardless of age or intellectual skills. This, however, would require substantial verification via repeated experiments (using the spoken mode) aimed at determining

whether the ceiling effects would hold for lower intellectual profile NSs (Hulstijn, 2012).

Concerning the NNSs, who showed fairly consistent variation on all MC tests, the open question as to whether or not late L2ers such as these can fully acquire BLC (Hulstijn, 2011) predicts that if L2ers can acquire any of the skills measured in the MC Test Battery to a nativelike level, this would be most difficult for Test 1-Phrasal Verbs-R and –P, Test 5-Topic Transition-R and Test 7-Feelings. This prediction meets an interesting set of results when statistical differences between areas of L1 and L2 metaphoric competence estimated in the next chapter.

Conversely, the fact that other tests (particularly Test 4-Topic/Vehicle-P and Test 8-Idiom Extension-R) elicited some level of NS variation implies that in most cases, the MC Test Battery involved more prototypically HLC tasks. To the extent that this is true, it bodes well for L2 learners, who (depending on intellectual skills, education, professional careers and leisure-time activities) can acquire HLC in their L2 to the NS level (Hulstijn, 2011).

6.4 Chapter summary

In this chapter, the key findings of **Analysis 1: The development and reliability of the MC Test Battery, descriptive statistics**, were discussed in relation to previous metaphoric competence and other research. Concerning the first research question, the extent to which MC can be reliability elicited and measured was complicated by several issues. The first set of problems concerned mixed levels of (particularly instrument) reliability. Secondly, test development, administration and refinement issues were discussed. These included problems eliciting metaphor, likely test taker anxiety and forming an orderly approach to data cleaning. Thirdly, high levels of NS variation for tasks involving acceptability of Vehicle extensions problematised Low's (1988) discussion by showing that a reliable NS base for judging NNS knowledge against (if this is what one wants to do) cannot always be ascertained. A final set of problems concerned the applicability of response data obtained from receptive multiple-choice questions to the real-world and to pedagogy. This highlighted the fact that these findings (and those in further chapters) are limited to the way in which MC was tested and the participant samples.

In spite of these problems, the discussion also emphasised how the decisions taken helped alleviate some of the problems of measuring metaphoric competence. Two main points can be made. First, while substantial amounts of data cleaning seem to be a 'necessary evil' in metaphoric competence test design and do not always lead to statistically reliable instruments (Littlemore, 2001), the process of removing problematic tests, items and participants is likely to have made data a truer representation of the constructs targeted. Second, the attention given to task instructions, examples, explanations and ordering of tests from receptive to productive and items from easy to difficult are argued to have had a facilitating and motivational effect for test takers. This can be further improved in future administrations of the MC Test Battery, since

researchers can use the reduced set of items resulting from data cleaning to save time and further mitigate test taker anxiety. The descriptive statistics also provide a guide as to which MC tests are apparently easier or more difficult than others, which research might use to revise the order in which MC tests are administered.

Concerning the second research question, apparent NNS and NS differences in MC test scores were discussed in terms of whether basic expectations were met, non-response patterns and the observed NS ceiling effects and variation in relation to Hulstijn's (2011, 2012) theory of basic and higher language cognition. This discussion highlighted that although Low (1988) and Littlemore and Low (2006a) left few clues as to the comparative ease and difficulty of their metaphor-related skills and (sub)competences for NNSs, two areas that were expected to be difficult for the NNSs, phrasal verbs and idiom extension, were confirmed to be so. From an examination of the NNS and NS non-response rates for Test 8-Idiom Extension-P, specific idioms that elicited higher amounts of non-response were identified. Whereas Littlemore and Low's (2006a) finding in this area is limited to an unspecified number of 'advanced' learners adapting six idioms and the general outcome that most chose not to engage in the task, this study took a more robust approach and found that learners *can*, given ample support, extend idioms, and that some idioms are more (or less) extendable than others. A comparison of these patterns for the NNSs and NSs suggested that further studies should explore whether idioms that test takers are reluctant to extend tend to be the same for NNSs and NSs (as the results showed), or whether fundamental differences can be observed.

Finally, it was tentatively suggested that the NS ceiling effects for Test 1-Phrasal Verbs-R and –P, Test 5-Topic Transition-R and Test 7-Feelings R, and variation for other tests may indicate areas of metaphoric competence that pertain to basic and higher levels cognition respectively. The implication for L2-ers is that these may be the most difficult aspects of MC to fully acquire, if this is at all possible. These issues require further exploration using the spoken mode and NSs or lower intellectual profiles.

Thus far, findings on NNS and NS differences have concerned superficial descriptive statistics. In the next chapter, MC and vocabulary knowledge and L2 proficiency are modelled to find latent underlying variables. NNS and NS differences between latent variables derived from the MC and vocabulary test data are then sought.

Chapter 7: Analysis 2 - Metaphoric and other (sub)competences uncovered

7.1 Introduction

In the first part of this chapter, an Exploratory Factor Analysis (EFA) of the NNS data is presented. The goal of EFA is to achieve parsimony via the generation of a model that explains the maximum amount of common variance in a correlation matrix using the smallest number of explanatory concepts (A Field, 2013; Tinsley & Tinsley, 1987). In EFA, factors are estimated using a mathematical model which analysis only the variable shared between variables. This model allows the observed data (i.e., the actual test scores) to be expressed as functions of a smaller number of possible causes.[62] The results of the EFA of the NNS data (NNS EFA) answer research question three, which asked about the extent to which factors underlie the observed L2 metaphoric competence, vocabulary knowledge and proficiency test scores, and what kinds of (sub)competences these factors might represent. In order not to impose prior assumptions about how L2 metaphoric competence, vocabulary knowledge and general proficiency relate, all 23 tests[63] were submitted for factor analysis rather than MC tests alone.

The second part of the chapter reports another EFA, this time on the NNS+NS data combined.[64] This analysis answered the first part of research question four, which asked about the extent to which the same factors can be found in the NNS and combined NNS+NS data. (see below, this section, for rationale).

In the third part of the chapter, Multivariate Analysis of Variance (MANOVA) and independent-samples *t*-tests were used to discover any statistical differences between the NNS and the NSs on both a combination of all factors (i.e., dependent variables), and for factors individually. These results answered the second half of research question four, which asked how NNS and NS factors scores differed. At this point, the reader may rightly ask *why not conduct separate EFAs of the NNS and NS data and compare those?* The reason why not is that even if the NS sample size were sufficient for EFA (it was not), the variables contributing to factors (and calculation of factor scores) in the two solutions would be different. Consequently, Multivariate

[62] In conventional factor analysis terminology, underlying skills or traits (factors) are said to 'affect' or 'cause' scores on observed tests, although in the exploratory case, causality is hypothesised rather than confirmed.

[63] Test 1-Phrasal Verbs-R and –P, Test 2-Metaphor Layering-R, Test 3-Vehicle Acceptability-R, Test 4-Topic/Vehicle-R, Test 5-Topic Transition-R and –P, Test 6-Heuristic-R and -P, Test 7-Feelings-R and -P, Test 8-Idiom Extension-R and -P, Test 9-Metaphor Continuation-R and –P, VYesNo, Word Associates Test, OOPT Use of English, OOPT Listening, IELTS Reading, IELTS Writing, IELTS Speaking, and IELTS Listening.

[64] Unlike the NNS EFA, which included all MC tests (except 4P), vocabulary tests and proficiency components, only the MC and vocabulary tests were submitted as variables in the EFA of the NNS+NS data because the NSs had not completed the OOPT and IELTS strands.

Analysis of Variance (MANOVA), which identifies statistical differences between groups on two or more dependent variables, could not be implemented. The only way to obtain directly comparable NNS and NS factor scores, and thus answer the second part of research question four, was to treat the NNSs and NSs as an intact sample for EFA. The factor scores generated were then submitted as dependent variables in a MANOVA, and ANOVAs and independent samples *t*-tests, with L1 group (i.e., NNS or NS) as the independent variable, conducted. Put simply, the analyses in part three showed whether or not there was any statistical difference in the metaphoric competence of the NNS and NS groups, and if so, to what extent and for which aspects (i.e., factors) of metaphoric competence.

Factor analysing NNS and NS data together may seem strange given that these two groups are usually treated as separate populations in L2 research.[65] To the best of the author's knowledge, there is no comparable study in the metaphoric competence literature, however, the legitimacy of factor analysing NNSs and NSs together is confirmed by Field (2013, pp. 673-674), who describes an equivalent scenario in which a *t*-test on factor scores for sociability obtained from an EFA on mixed-sex subjects might be used to discover whether females are significantly more sociable than males. Field's males and females, who constitute two populations of interest, are comparable with the NNSs and NSs in the present study, who are factor analysed together and then treated as two groups of a categorical independent variable for MANOVA.

7.2 EFA of NNS data: Discovering underlying L2 metaphoric (sub)competences

7.2.1 Data screening

Since EFA involves several important assumptions, the NNS data were first screened using Tabachnick and Fidell's checklist (2013, p. 125). Presented below is a summary of this process, supplementary tables and figures can be found in Appendix G. Two things to note are that all decisions pertain to criteria stipulated by the authors unless otherwise indicated, and the term 'variables' is used throughout to mean the 23 tests submitted for analysis.

Data screening revealed that:

- Missing data (i.e., deleted scores) were missing completely at random and comprised less than 1% spread across multiple variables, suggesting unproblematic randomness rather than issues with specific tests;

[65] By factor analysing the NNS+NS data together, I (of course) do not intend to claim anything about what the factor structures would look like if the two groups could be analysed separately.

- 'Pairwise' deletion, where missing scores are removed only from the variable in question (not the participant's scores for other variables), was the best approach to missing data since it offered more statistical power than 'listwise' deletion, and is not as controversial as the 'replace with mean' method (Pallant, 2013);
- Most MC and all IELTS variables were nonnormally distributed but vocabulary and OOPT variables were normally distributed, evidenced numerically via Kolmogorov-Smirnov test, Shapiro-Wilk test (p < .01 indicating nonnormality), skewness and kurtosis measures, and visually via histograms and Q-Q plots;
- Data met assumptions of linearity evidenced by an absence of curvilinearity in scatterplots of the most discrepant positively and negatively skewed variables;
- Although some heteroscedasticity was found, this was unproblematic given that data were not curvilinear;
- No participants were multivariate outliers, evidenced by the fact that all Mahalanobis distances were below the cut-off value of 49.728, the critical value of chi-square when the degree of freedom = 23 [variables] (p < .001);[66]
- Data were multivariate nonnormal, evidenced by the fact that not all variables were univariate normal, a necessary (but insufficient) condition of multivariate normality, the results of Mardia's, Henze-Zirkler's and Royston's tests, and inspection of chi-square Q-Q plots;[67]
- Data met the assumptions for multicollinearity, evidenced by the absence of correlations above .80 (A Field, 2013), the fact that all variance inflation factors (VIF) were below the upper limit of 10 and all tolerance statistics above 0.1 (a serious problem) and 0.2 (a potential problem) (A Field, 2013; Pallant, 2013),[68] and the finding from the collinearity diagnostics that none of the six roots (dimensions) with condition indexes above 30 were coupled with variance proportions greater than .50 for two or more different variables (Tabachnick & Fidell, 2013);
- The sample size of 112 (110 for some variable pairs) and sample-to-variable ratio of 4.78 were borderline adequate by a 'recommended absolute minimum' approach (Hair, Anderson, Tatham, & Black, 1995; Maccallum, Widaman, Zhang, & Hong, 1999; Pallant,

[66] More robust multivariate outlier detection methods than inspection of Mahalanobis distances are available, but were not used because their accuracy diminishes with smaller samples like that of the present study (Hardin & Rocke, 2005).

[67] Conducted using a web-tool application based on an MVN package from R (Korkmaz, Goksuluk, & Zararsiz, 2014), publicly available at http://www.biosoft.hacettepe.edu.tr/MVN/. These tests and plots were inconclusive: Mardia's and Royston's tests indicated multivariate nonnormality, whereas Henze-Zirkler's showed multivariate normality. Neither was a comparison of the plots with the authors' example plots very illuminating. However, given the requirement of univariate normality of all variables, data were most probably multivariate nonnormal.

[68] The closest to a 'problematic' variable was the VYesNo (VIF = 2.634, tolerance statistic = .380).

2013; Tabachnick & Fidell, 2013), about half way between the minimum and median sample sizes of 25 and 253 respectively for EFAs in published SLA research (Plonsky & Gonulal, 2015), less than half the recommended sample-to-variable of 10-15 participants per variable (A Field, 2013),[69] and 'good' given that the overall Kaiser-Meyer-Olkin (KMO) test statistic for the factor model specified was .84 and above .5 for all individual variables (A. Field, Miles, & Field, 2012);

- The best solution to the problem of the small sample size and uni- and multivariate nonnormality was to use Principal Axis Factoring (PAF), a 'descriptive' factor extraction method[70] that "has the advantage of entailing no distributional assumptions" (Fabrigar, Wegener, MacCallum, & Strahan, 1999, p. 277),[71] in combination with bootstrapping procedures, which treat the sample as the population and repeatedly extract thousands of smaller random samples from this, to increase statistical power (LaFlair, Egbert, & Plonsky, 2015; Plonsky, Egbert, & LaFlair, 2015), help determine the number of factors to retain, and as an internal method of estimating the replicability of pattern coefficients over various resamples without the need to assume normality (T. J. B. Kline, 2005; Zientel & Thompson, 2007).[72] Transformations of variables was explored, but did not improve normality and minimised skewness at the expense of increasing kurtosis (and vice-versa) and so was not implemented;
- Data met the first two basic assumptions of EFA, namely that each of the variables consisted of scale data, and variables were linearly related and moderately correlated.

7.2.2 Factor retention

Given the complex nature of EFA, decisions about the number of factors to retain based on multiple criteria rather than one method are strongly encouraged, although this has not been standard practice in SLA research (Brown, 2009; Loewen & Gonulal, 2015; Plonsky & Gonulal, 2015). Plonsky and Gonulal (2015), for instance, found that in 73.2% of studies analysed (K = 37), either a single criterion was used, or factor retention criteria were not reported at all.

In response to this shortcoming, several methods for determining the number of factors

[69] The sample-to-variable ratio of 4.78 was in fact within both the range of previously proposed minimum ratios (3 to 20) and actual ratios observed (3 to 76) in SLA research (Plonsky & Gonulal, 2015).

[70] 'Descriptive' factor extraction methods (also including Principle Components Analysis and Image Factoring) assume that the sample used is the population. Thus, findings generalise to the actual wider population only if the factor structure can be replicated using a different sample (A Field, 2013; Tinsley & Tinsley, 1987).

[71] Tabachnick and Fidell comment that "as long as PCA and EFA are used descriptively as convenient ways to summarize the relationships in a large set of observed variables, assumptions regarding the distribution of variables are not in force" (2013, p. 666).

[72] Bootstrapping was implemented using R, via routines developed by Zopuoglu (2017a, 2017b) publicly available at https://sites.education.miami.edu/zopluoglu/software-programs/.

were used:

Table 7.1 *Criteria for Retaining Factors*

Method	Criterion	Description
1	Kaiser's > 1 rule	Retain factors with eigenvalues greater than one (Guttman, 1954; Kaiser, 1960)
2	Bootstrap lower 95% CI with Kaiser's > 1 rule[73]	Same as method 1, but uses lower bound 95% confidence intervals from 5,000 resamples rather than initial eigenvalues
3	Joliffe's > 0.7 rule	Retain factors with eigenvalues greater than 0.7 (Joliffe, 1972, 1986)
4	Boots. lower 95% CI with Joliffe's > 0.7 rule	Same as method 3, but uses lower bound 95% confidence intervals from 5,000 resamples rather than initial eigenvalues
5	Scree plot: Point(s) of inflexion	Retain factors to the left of the point of inflexion in scree plot of eigenvalues against factors (Cattell, 1966)
6	Parallel analysis (in R)[74]	Retain factors with eigenvalues lower than randomly generated counterparts (Horn, 1965) as shown in R
7	Parallel analysis (in SPSS)[a]	Same as method 6, but use SPSS instead of R
8	Total variance explained > 20%	Retain factors to exceed a minimum of 20% total variance explained (Brown, 2009 in combination with SLA field minimum found by Plonsky and Gonulal, 2015)

[a] Rejecting initial eigenvalues lower than 95th percentile, both random data and raw data permutation methods tested.

Individually, each of these methods have advantages and disadvantages. Kaiser's criterion is accurate when the number of variables is less than 30 and all commonalities after extraction exceed 0.7, or when the sample size exceeds 250 and the average communality is greater than .6 (A Field, 2013), whereas Joliffe (1972), using 587 sets of randomly generated data, found satisfactory results when factors with eigenvalues greater than 0.7 were retained.[75] Bootstrapping and parallel analysis are advantageous because they give the decision an empirical basis, but can be criticised on theoretical grounds, precisely because they employ randomly generated data sets (Green, Levy, Thompson, Lu, & Lo, 2012; Harshman & Reddon, 1983; Ruscio & Roche, 2012; Turner, 1998). The Scree plot can be difficult to interpret, but is thought to be reliable when sample size exceeds 200 (A Field, 2013). And while a minimum of 20% total variance explained is easily implementable, a solution with this amount of explanatory power is questionable, but not necessarily worthless (Plonsky & Gonulal, 2015).

Taken together, the various criteria suggest retaining anywhere from 1 to 12 factors, with little agreement between methods (Table 7.2).

[73] Zopuoglu's (2017a, 2017b) syntax does not appear to be amenable to 'pairwise' deletion, so the more controversial 'replace with mean' approach to missing data was used, although this is not likely to have resulted in anything except negligible differences.

[74] Traditional parallel analyses in R was conducted using the 'fa.parallel' function with the psych package, and in SPSS using rawpar.sps syntax developed by Brian O'Connor (2000), publicly available at https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html

[75] Since an eigenvalue of 1 explains as much variance as a variable, those in excess of 1 offer comparatively more explanatory power, and there is little to recommend Joliffe's criterion over Kaiser's unless it produces a factor structure that is better supported by better theoretical evidence (Cortina, 2002).

Table 7.2 *Number of Factors to Retain by Multiple Criteria*

Criterion	Suggested number of factors to retain
Kaiser's > 1 rule	7
Bootstrap lower 95% CI with Kaiser's > 1 rule	6
Joliffe's > 0.7 rule	12
Boots. lower 95% CI with Joliffe's > 0.7 rule	10
Scree plot: Point(s) of inflexion	1
Parallel analysis (in R)	2
Parallel analysis (in SPSS)[a]	1
Total variance explained > 20%	1 or more

[a] initial eigenvalues lower than 95th percentile rejected, both random data and raw data permutation methods tested.

As expected, stricter methods such as the scree plot and parallel analysis suggest the fewest number of factors, whereas the most liberal method, Joliffe's greater than 0.7 rule, suggests the most. Guiding the decision was the consideration that a good solution is one that is parsimonious yet explains a reasonable amount of variance in the data, yields several well defined factors, and not least of all "makes sense" in terms of its interpretability (Tabachnick & Fidell, 2013, p. 661). The six-factor solution, suggested by the bootstrap lower 95% CI with Kaiser's rule, was the best overall match for these criteria. The advantages of the six-factor solution were that it explained 17% more total variance than a one-factor solution, yielded interpretable factors, was empirically grounded (via the bootstrap eigenvalues), and constituted a mid-way point between the extremities of one and 12 factors. The appropriateness of the proposed six-factor solution was evaluated, and further supported via post hoc assessment of the model (Appendix H). The implications of this decision and possible future research into competing models are covered in subsequent discussion chapters.

7.2.3 Factor rotation

In EFA, variables characteristically have high loadings on the most important factor, and low loadings on all others. Consequently, interpretation of loadings can be difficult and so a 'rotation' is performed to help discriminate between factors without changing the underlying mathematical properties of the model (A Field, 2013; Tabachnick & Fidell, 2013). Whereas orthogonal rotations assume that factors are uncorrelated, oblique rotations permit correlated factors, which are to be expected where factors are connected to human cognition (A Field, 2013; Plonsky & Gonulal, 2015). Since metaphoric competence, L2 vocabulary knowledge and L2 language proficiency are all highly related to cognition and underlying competencies affecting scores on the administered tests would not be expected to be completely independent, factors were assumed to be correlated. Therefore, direct oblimin, a common and recommended method of oblique rotation was used. The NNS data were then factor analysed.

7.2.4 Results

7.2.4.1 Factor structure

Results from both SPSS and R show that the six-factor solution explained 42% total variance (cumulative percentage) by the extraction sums of squared loadings. Despite differences in the statistics presented and distributions of percentage explained by each factor, the fact that the total variance (cumulative percentage) explained is the same in both the SPSS and R data suggests that the model is reliable across software programmes.[76] The SPSS solution shows that before extraction, the first factor explained 27.79% of the total variance; this amount diminishes quite rapidly for factors 2, 3, 4, 5 and 6. In R, an evenly distributed percentage of variance was explained by each extracted factor: (F1) 8.39%, (F2) 8.35%, (F3) 6.57%, (F4) 5.87%, (F5) 6.26%, (F6) 6.74%. Notably, although a general decrease is observed, factors 5 and 6 explained slightly more variance than some previous factors.

With oblique rotation, the factor matrix is split into two matrices: the structure matrix and pattern matrix. The structure matrix (Appendix H) contains the correlation coefficients between each variable and each factor, however, since factors correlate these coefficients are most likely inflated by overlap between factors, i.e., variables may correlate with factors through a factor's correlation with another factor rather than directly (Tabachnick & Fidell, 2013). The pattern matrix, on the other hand, contains regression coefficients for each variable on each factor and shows the unique contribution of a variable to each factor, making it the favoured choice for interpreting the solution (A Field, 2013). For these reasons, the structure matrix is included in Appendix H and the interpretation of the pattern matrix is presented here.

The pattern matrix below (Table 7.3) shows the size of each variable's unique loading on each of the six factors. Only 'substantial' loadings of above 0.30 are shown. Colours are provided here and throughout for ease of factor recognition. Four variables did not load substantially on any of the factors and so no loadings are shown for these. Factor 1 was defined by a high loading for VYesNo (0.86) and IELTS Writing (0.40). Factor 2 seems to affect IELTS variables, as well as two MC test. Factor 3 is defined by four variables, with Test 1-Phrasal Verbs-P (0.62) as the best marker. Factor 4 is defined by two productive MC tests

[76] Differences are either due to the fact that the psych.fa package wrongly computes sums of squared Loadings for oblique rotation (Žiberna, 2015), or because both programmes use alternative rotation algorithms, as pointed out with regard to a similar problem by StackExchange blogger 'ttnphns' (ttnphns, 2012), a Russian statistician.

Table 7.3 *Pattern Matrix NNS EFA*

Test/Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Communality	Uniqueness	Complexity
IELTS Listening		0.62					0.58	0.42	1.3
IELTS Reading		0.61					0.49	0.51	1.2
Test 2-Metaphor Layering-R		0.41					0.40	0.60	2.4
IELTS Speaking		0.32					0.35	0.65	3.3
Test 6-Heuristic-R		0.30					0.20	0.80	2.4
VYesNo	0.86						0.82	0.18	1.0
IELTS Writing	0.40						0.42	0.58	3.6
Test 9-Metaphor Continuation-R							0.21	0.79	3.9
Test 1-Phrasal Verbs-P			0.62				0.42	0.58	1.2
OOPT Listening			0.49				0.50	0.50	2.0
Test 1-Phrasal Verbs-R			0.41				0.24	0.76	2.2
OOPT Use of English			0.36				0.46	0.54	2.7
Word Associates Test							0.55	0.45	4.5
Test 8-Idiom Extension-P						0.76	0.68	0.32	1.0
Test 9-Metaphor Continuation-P						0.37	0.49	0.51	2.9
Test 8-Idiom Extension-R						0.33	0.25	0.75	2.4
Test 7-Feelings-R							0.35	0.65	3.8
Test 4-Topic/Vehicle-R					0.59		0.32	0.68	1.2
Test 5-Topic Transition-R					0.53		0.34	0.66	1.2
Test 3-Vehicle Acceptability-R					0.48		0.30	0.70	1.4
Test 6-Heuristic-P				0.81			0.67	0.33	1.0
Test 7-Feelings-P				0.32			0.29	0.71	3.2
Test 5-Topic Transition-P							0.36	0.64	4.8

measuring illocutionary skills. Factor 5 seems to concern Topic and Vehicles, and factor 6 is defined by idiom extension and metaphor continuation tests. The interpretation of factors was conducted systematically, and is reported in section 7.3.4.2.

Using Comrey and Lee's (1992) guidelines for assessing the extent to which a variable is a pure measure of a factor, factors 1, 4 and 6 are marked by 'excellent' measures (> 0.71, indicating 50% overlapping variance), and factors 2, 3 and 5 are marked by 'good' measures (around 0.55, 30% overlapping variance). Using Stevens' (2002) rule-of-thumb that with sample sizes in excess of 100 participants loadings over 0.512 are likely to be statistically significant, eight variables were likely to have had significant loadings.

The communalities column shows the proportion of each variable's variance explained by the extracted factors (A Field, 2013), the uniqueness is essentially the inverse of this and is calculated as 1 minus the communality. VYesNo, Test 8-Idiom Extension-P, and Test 6-Heuristic-P had the highest communalities, indicating that the six factor solution accounted for large amounts of variance in these variables. Low communalities (e.g., for Test 6-Heuristic-R) indicate that variables may not fit well with others on the same factors (Pallant, 2013).

Finally, each variable's score on Hofmann's (1978) row-complexity index is shown in the complexity column. These scores denote the average number of factors needed to account for the measured variable (Hofmann, 1978; Pettersson & Turkheimer, 2010). Lower scores are desirable, since they indicate a 'purer' variable in the sense that it is associated with a fewer number of factors (ideally one). The variables that did not load substantially on any factor have notably higher complexity scores, suggesting their association with several factors. Of the variables that did load substantially, IELTS Writing, IELTS Speaking and Test 7-Feelings-P were also found to be fairly complex variables while the purest variables were VYesNo, Test 8-Idiom Extension-P, and Test 6-Heuristic-P.

Since it was not possible to verify the pattern matrix loadings via replication on another sample of participants, bootstrapping techniques were used as a method of 'internally' estimating replicability via random resamples of the data (Zientel & Thompson, 2007). Table 7.4 shows the pattern matrix loadings from the original sample, and mean bootstrap loadings, standard deviations (i.e., estimated standard errors), the coefficient of variation (CV) from these statistics,[77] and lower and upper bound 95% confidence intervals from 5,000 resamples. In order to resolve the problem that factors may vary in their order across resamples, Procrustes rotation was used to ensure a common factor space by rotating all resamples to a best-fit position (Zientel & Thompson, 2007).

The bootstrap mean loadings are, on average, slightly lower than the sample estimates

[77] CVs are provided to enable comparison of the spread of data with other studies.

Table 7.4 *Bootstrapping of NNS EFA Pattern Matrix Loadings across 5,000 Resamples*

			Bootstrap estimate				
						95% Confidence interval[a]	
Test/Variable	Factor	Pattern matrix loading	*M*	*SD*	*CV*[b]	Lower	Upper
VYesNo	1	0.86	0.64	0.16	0.25	0.55	1.18
IELTS Writing	1	0.40	0.44	0.14	0.32	0.12	0.67
IELTS Listening	2	0.62	0.56	0.12	0.21	0.39	0.85
IELTS Reading	2	0.61	0.54	0.12	0.22	0.38	0.84
T2-Metaphor Layering-R	2	0.41	0.42	0.11	0.27	0.19	0.63
IELTS Speaking	2	0.32	0.34	0.12	0.35	0.09	0.55
T6-Heuristic-R	2	0.30	0.29	0.15	0.51	0.01	0.59
T1-Phrasal Verbs-P	3	0.62	0.58	0.13	0.22	0.37	0.86
OOPT Listening	3	0.49	0.45	0.14	0.31	0.23	0.76
T1-Phrasal Verbs-R	3	0.41	0.40	0.14	0.35	0.13	0.68
OOPT Use of English	3	0.36	0.36	0.12	0.32	0.13	0.59
T6-Heuristic-P	4	0.81	0.61	0.17	0.27	0.49	1.13
T7-Feelings-P	4	0.32	0.33	0.16	0.47	0.01	0.63
T4-Topic/Vehicle-R	5	0.59	0.49	0.19	0.38	0.22	0.95
T5-Topic Transition-R	5	0.53	0.48	0.18	0.37	0.18	0.87
T3-Vehicle Acceptability-R	5	0.48	0.43	0.15	0.35	0.19	0.77
T8-Idiom Extension-P	6	0.76	0.59	0.15	0.24	0.48	1.05
T9-Metaphor Continuation-P	6	0.37	0.39	0.13	0.33	0.11	0.62
T8-Idiom Extension-R	6	0.33	0.37	0.16	0.45	0.01	0.65

[a] 95% confidence intervals were calculated as the pattern matrix loading minus (for lower bound) or plus (for upper bound) 1.96 times the Standard Error (i.e., the SD of bootstrap estimates).

[b] *CV* = Coefficient of variance, a standardised measure of spread calculated as the Standard Deviation divided by the Mean bootstrap estimate.

(on average, around 0.46 compared with 0.50). The standard deviations of the bootstrap estimates (i.e., the standard errors) and coefficients of variation revealed that factor 2 contained variables with both the highest and lowest dispersions across 5000 resamples (Test 6-Heuristic-R and IELTS Listening respectively). Factor 1 had, on average, the least dispersed variables and hence was most stable across resamples.

The confidence intervals provide the lower and upper bounds within which the true (i.e., population) value of a loading coefficient is likely to lie, with 95% probability, and thus provide useful information about the generalisability of findings and likelihood of replicability. Most of these estimates are quite wide, indicating that the observed structure and strength of loadings obtained from the participants in the present study may not replicate exactly with different pools of participants from the same population. This is somewhat unsurprising, especially given the fact that in EFA, factor structures can differ (i.e., not replicate) with even slight variations in method of factor retention, deletion of cases, variables submitted for analysis, factor extraction method, rotation method and other methodological steps (Tabachnick & Fidell, 2013). Notwithstanding, the confidence intervals do reveal that in 95% of 5,000 bootstrap resamples (i.e., replications sampling from the dataset), the strongest marker variables for each factor (except factor 5) loaded substantially (i.e., > 0.30) on the same factors.

7.2.4.2 Interpretation of factor loadings

The statistical adequacy of the six-factor model was confirmed by several indicators (Appendix H). Another important consideration in evaluating the adequacy of an EFA is its interpretability: "a good PCA or FA 'makes sense'; a bad one does not" (Tabachnick & Fidell, 2013, p. 661). The interpretation of factors is the subject of this section.

The process of interpreting factors is sometimes regarded as something of an art. This is probably due to the absence of hard and fast rules and the need for creativity yet precision when assigning names to factors. Nevertheless, it is both possible and important to be principled. Therefore, in order to interpret factors in the present study, information about the strength of each variable-to-factor loading, loading stability defined as whether or not the lower 95% confidence interval for 5,000 bootstrap resamples indicated a substantial loading (> 0.3), and detailed descriptions of what each variable measured were compiled. In this way, it was possible to give more interpretive weight to stronger and more stable markers (Tabachnick & Fidell, 2013). This information is presented in Table 7.5.

Table 7.5 *Information for Interpreting Factors in the NNS EFA*

Test/Variable	Factor	Load.	Strength of load.	Population stability (low 95% CI > 0.3)	Description of what each variable measured
VYesNo	1	0.86	Excellent	Y	vocabulary size from 0-10k words known using the Yes/No format
IELTS Writing	1	0.40	Poor	N	task achievement, coherence and cohesion, lexical resource, grammatical range and accuracy in writing
IELTS Listening	2	0.62	Good	Y	ability to understand main ideas, facts, opinions, attitudes, purposes, and follow arguments when listening
IELTS Reading	2	0.61	Good	Y	ability to understand main ideas, details, implied meanings, opinions, attitudes and follow arguments in texts
T2-Metaphor Layering-R	2	0.41	Poor	N	ability to understand layers of figurative and literal meaning in metaphors and puns
IELTS Speaking	2	0.32	Poor	N	task achievement, coherence and cohesion, lexical resource, grammatical range and accuracy in speaking
T6-Heuristic-R	2	0.30	Poor	N	ability to recognise similes used to perform heuristic functions
T1-Phrasal Verbs-P	3	0.62	Good	Y	ability to recall a metaphorical phrasal verb particle
OOPT Listening	3	0.49	Fair	N	ability to identify the literal, intended, and implied meanings being communicated in what is heard
T1-Phrasal Verbs-R	3	0.41	Poor	N	ability to recognise a metaphorical phrasal verb particle
OOPT Use of English	3	0.36	Poor	N	grammatical and pragmatic knowledge of English
T6-Heuristic-P	4	0.81	Excellent	Y	ability to recall similes to perform heuristic functions
T7-Feelings-P	4	0.32	Poor	N	ability to recall metaphors that convey information and feelings about that information
T4-Topic/Vehicle-R	5	0.59	Good	N	ability to rate the acceptability of different Topic and Vehicle combinations in the framework of an analogy
T5-Topic Transition-R	5	0.53	Fair	N	ability to recognise proverb/idioms in topic transition in interactive discourse
T3-Vehicle Acceptability-R	5	0.48	Fair	N	ability to rate the acceptability of Vehicle terms across different word classes
T8-Idiom Extension-P	6	0.76	Excellent	Y	ability to create possible extensions of idioms
T9-Metaphor Continuation-P	6	0.37	Poor	N	ability to create coherent continuations of metaphoric discourse
T8-Idiom Extension-R	6	0.33	Poor	N	ability to recognise possible extensions of idioms

Factor 1, defined by VYesNo as a strong and stable marker, appeared to suggest the construct of **English Vocabulary Size**. Curiously, in this analysis, the Word Associates Test did not load substantially onto this or any other factor as might have been expected. I consider this in the next chapter. Although it emerged as its own factor, vocabulary size is likely to have played a role in all tests, given that all were linguistic (rather than non-linguistic) measures. IELTS Writing also loaded poorly and non-stably on this factor. To the extent that this was non-coincidental, it can probably be explained by the influence of the 'lexical resources' scoring component on this particular IELTS strand.

In order to measure the reliability of NNS factors, in terms of both the internal consistency (or correlation if only two tests) of tests-within-factors and (test) items-within-factors, Cronbach's alpha and correlation coefficients were calculated. Results of this analysis be found in Appendix H, and are discussed for each factor in turn. For **English Vocabulary Size**, VYesNo and IELTS had a 'medium' strength correlation,[78] significant at the .01 level (r = 0.47, n = 110), suggesting some conceptual relatedness of these tests.

Factor 2 was marked most strongly and stably by IELTS Listening and IELTS Reading, but also poorly and non-stably by Test 2-Metaphor Layering-R, IELTS Speaking and Test 6-Heuristic-R. Taking the strongest markers as a guide, this factor was labelled **English General Comprehension**, a competence that could conceptually connect the two 'receptive' IELTS skills and the two metaphoric competence tests. The fact that IELTS Speaking (i.e., a productive skill) loaded onto this factor poorly is probably reflective of its stronger correlation with IELTS Listening and Reading than Writing. This may also suggest that NNSs who had better comprehension skills in both the spoken and written modes, were also better L2 speakers. Sticking with the adjectives selected to discuss Cronbach's alpha in the present study (section 6.2.1), the internal consistency of **English General Comprehension** tests-within-factor was 'low' (α = 0.310), while the mean internal consistency of items-within-factor[79] was 'moderate' (α = 0.620).

Factor 3 was most strongly defined by Test 1-Phrasal Verbs-P, but also by Test 1-Phrasal Verbs-R and the two components of the OOPT. The OOPT's Use of English section measures (alongside other areas of grammar and pragmatics) knowledge of the forms, particles, and separability of phrasal verbs (Purpura, 2009). The listening section of the OOPT measures how well test takers can apply this knowledge during listening. This may explain why these tests loaded with the phrasal verbs tests from the MC Test Battery. Cumulatively, a common theme of the four tests appears to be grammar and structures, with phrasal verbs (particularly recalling

[78] As in previous chapters, I use Pallant's (2013, p. 139) guidelines as convenient ways to refer to correlation strength: r = .10 to .29 (small); r = .30 to .49 (medium); r =- .50 to 1.0 (large).
[79] Because groups 1 and 2 had completed different versions of the test, a mean of both group's alphas was taken.

the *correct* particle) serving as something of an indicator of this. This factor was therefore called **English Grammatical Metaphoric Competence**. The internal consistency of **English Grammatical Metaphoric Competence** was 'moderate' (α = 0.602) for tests-within-factor and 'low' for items-within-factor (α = 0.460).

Factor 4 was named **English Illocutionary Metaphor Production** on account of being marked by two productive MC tests from within the illocutionary dimension of Littlemore and Low's (2006a) description of metaphoric competence. Both of these tests required participants to produce a simile or metaphor for particular illocutionary purposes; namely, explaining something to a child, and conveying feelings about something. Interestingly, both tests involved the elicitation of similes and thus bear a syntactic as well as illocutionary connection. The tests loading on **English Illocutionary Metaphor Production** displayed a 'medium' strength, positive correlation significant at the .01 level (r = 0.34, N = 112), and 'moderate' internal consistency of items-within-factor (α = 0.584).

Factor 5 is characterised by good and fair (though non-stable) loadings on three receptive MC tests, which all involved Topics and/or Vehicles. The fact that both Test 3-Vehicle Acceptability-R and Test 4-Topic/Vehicle-R used the empirical NS data to score NNS ratings probably also contributed to their loading on the same factor. This factor was labelled **English Topic/Vehicle Acceptability** because it seems to primarily concern the ability to rate the acceptability of various Vehicles from the same domain, or Vehicle terms on their own. The internal consistency of tests-within-factor was 'moderate' (α = 0.559), and 'high' for items-within-factor (α = 0.729).

Finally, **factor 6** was marked strongly and stably by Test 8-Idiom Extension-P, and poorly and non-stably by Test 9-Metaphor Continuation-P and Test 8-Idiom Extension-R. These tests all involve a high degree of creativity, novelty and language play. Therefore, the factor was named **English Metaphor Language Play**. This factor displayed 'moderate' level of internal consistency for test-within-factor (α = 0.677), but 'very high' internal consistency for items-within-factor (α = 0.834).

The factor correlation matrix is contained in Appendix H and shows that the strongest correlation was between **(F1) English Vocabulary Size** and **(F2) English General Comprehension** (r = .48) and the lowest was between **(F2) English General Comprehension** and **(F4) English Illocutionary Production** (r = .20). The factor structure, in comparison with past research, is discussed in the next chapter.

7.3 EFA of NNS+NS data: Do the same factors appear when all data are analysed together?

7.3.1 Data screening

Unlike the NNS EFA, which included all MC tests (except 4P), vocabulary tests and proficiency components, only the MC and vocabulary tests were of interest as variables in the NNS+NS EFA because the NSs had not completed the OOPT and IELTS strands. This meant that the NNS+NS EFA concerned 17 'variables' (i.e., tests). Data were screened in exactly the same way as for the NNS EFA (Tabachnick & Fidell, 2013, p. 125) and found to be suitable for EFA but again due to small sample size and nonnormality, requiring the use of bootstrapping and Principal Axis Factoring. Due to limited space, the various data screening analyses are not reported, but are available upon request.

7.3.2 Factor retention

Concerning the number of factors to retain, the same multiple criteria were used as with the NNS EFA. Bootstrap Kaiser's rule 95% lower CI, the Scree plot, parallel analyses and total variance explained suggested 1 factor, Kaiser's-greater-than-1 rule suggested 2 factors, bootstrap Joliffe's rule 95% lower confidence interval suggested 4 factors and Joliffe's rule suggested 5 factors. A four-factor solution was selected for two reasons: 1) two factors with eigenvalues less than one (but above .7) were interpretable and also found in the NNS solution, constituting sufficient theoretical grounds for their retention (Cortina, 2002); and 2) a one-factor solution, suggested by the stricter methods, explained around 10% less total variance.

7.3.3 Factor rotation

The rotation method selected was again direct oblimin, as factors were expected to be correlated.

7.3.4 Results

7.3.4.1 Factor structure

Since the NSs did not complete the L2 proficiency tests, only the MC Test Battery and vocabulary test variables were submitted for EFA. The pattern matrix and bootstrap loading of the four-factor model are presented below (Tables 7.6 and 7.7). With this solution, 61% of the total variance was explained after extraction (sums of squared loadings), substantially more than the 42% for the six factor solution in the NNS EFA. The statistical adequacy of the NNS+NS four-factor model was confirmed by several indicators.

Table 7.6 *Pattern Matrix NNS+NS EFA*

Test/Variable	Factor 1	Factor 2	Factor 3	Factor 4	Communality	Uniqueness	Complexity
T1-Phrasal Verbs-R	0.89				0.67	0.33	1.00
T2-Phrasal Verbs-P	0.85				0.72	0.28	1.00
T3-Vehicle Acceptability-R	0.72				0.76	0.24	1.30
T9-Metaphor Continuation-R	0.60				0.47	0.53	1.00
VYesNo	0.53				0.68	0.32	1.60
T7-Feelings-R	0.50				0.51	0.49	1.40
T6-Heuristic-R	0.49				0.39	0.61	1.50
Word Associates Test	0.46				0.72	0.28	2.10
T2-Metaphor Layering-R	0.43				0.71	0.29	2.30
T7-Feelings-P	0.38	0.31			0.43	0.57	2.00
T5-Topic Transition-P	0.33		0.31		0.68	0.32	3.00
T8-Idiom Extension-P			0.89		0.76	0.24	1.00
T9-Metaphor Continuation-P			0.60		0.65	0.35	1.30
T8-Idiom Extension-R			0.58		0.49	0.51	1.20
T6-Heuristic-P		0.90			0.81	0.19	1.00
T4-Topic/Vehicle-R				0.56	0.38	0.62	1.20
T5-Topic Transition-R				0.54	0.46	0.54	1.20

Table 7.7 *Bootstrapping of NNS+NS EFA Pattern Matrix Loadings across 5,000 Resamples*

			Bootstrap estimate				
						95% Confidence interval[a]	
Test/Variable	Factor	Pattern matrix loading	*M*	*SD*	*CV*[b]	Lower	Upper
T1- Phrasal Verbs-R	1	0.89	0.83	0.11	0.14	0.61	1.05
T1-Phrasal Verbs-P	1	0.85	0.78	0.10	0.13	0.58	0.97
T3-Vehicle Acceptability-R	1	0.72	0.70	0.10	0.14	0.51	0.90
T9-Metaphor Continuation-R	1	0.60	0.59	0.12	0.21	0.35	0.83
VYesNo	1	0.53	0.55	0.11	0.20	0.34	0.77
T7-Feelings-R	1	0.50	0.52	0.11	0.21	0.31	0.73
T6-Heuristic-R	1	0.49	0.48	0.13	0.26	0.23	0.72
Word Associates Test	1	0.46	0.51	0.09	0.18	0.33	0.69
T2-Metaphor Layering-R	1	0.43	0.47	0.09	0.19	0.29	0.64
T7-Feelings-P	1	0.38	0.38	0.13	0.33	0.14	0.63
T5-Topic Transition-P	1	0.33	0.39	0.10	0.26	0.19	0.59
T6-Heuristic-P	2	0.90	0.57	0.25	0.44	0.08	1.06
T7-Feelings-P	2	0.31	0.29	0.17	0.58	-0.04	0.62
T8-Idiom Extension-P	3	0.89	0.70	0.19	0.27	0.33	1.07
T9-Metaphor Continuation-P	3	0.60	0.53	0.13	0.25	0.26	0.79
T8-Idiom Extension-R	3	0.58	0.53	0.16	0.31	0.21	0.85
T5-Topic Transition-P	3	0.31	0.33	0.12	0.34	0.11	0.56
T4-Topic/Vehicle-R	4	0.56	0.35	0.30	0.84	-0.23	0.93
T5-Topic Transition-R	4	0.54	0.34	0.26	0.75	-0.16	0.85

[a] 95% confidence intervals were calculated as the pattern matrix loading minus (for lower bound) or plus (for upper bound) 1.96 times the Standard Error (i.e., the SD of bootstrap estimates).
[b] *CV* = Coefficient of variance, a standardised measure of spread calculated as the Standard Deviation divided by the Mean bootstrap estimate.

7.3.4.2 Interpretation of factor loadings

The same process as before was used to interpret factors. Table 7.8 contains information about the strength of each variable-to-factor loading, loading stability defined as whether or not the lower 95% confidence interval for 5,000 bootstrap resamples indicated a substantial loading (> 0.3), and detailed descriptions of what each variable measured.

The four factors identified in the NNS+NS data corresponded with four out of six found in the NNS data, especially concerning the strongest marker variables, which were the same for three of the NNS+NS factors and negligibly different for the remaining factor.[80] **NNS+NS Factor 1** was most strongly marked by the phrasal verbs tests and so was identified as **English Grammatical Metaphoric Competence** (a match of NNS factor 3). This factor was also strongly marked by Test 3-Vehicle Acceptability-R, which had phrasal verbs in some of its test items and (in part) measured sensitivity to grammatical acceptability.

[80] In the NNS EFA, Test 1-Phrasal Verbs-P was a strongest marker, whereas in the NNS+NS EFA, Test 1-Phrasal Verbs-R had a slightly higher loading than Test 1-Phrasal Verbs-P (0.89 compared to 0.85).

Table 7.8 *Information for Interpreting Factors in the NNS+NS EFA*

Test/Variable	Factor	Load	Strength of load.	Population stability (low 95% CI > 0.3)	Descriptions of what each variables measured
T1-Phrasal Verbs-R	1	0.89	Excellent	Y	ability to recognise a metaphorical phrasal verb particle
T1-Phrasal Verbs-P	1	0.85	Excellent	Y	ability to recall a metaphorical phrasal verb particle
T3-Vehicle Acceptability-R	1	0.72	Excellent	Y	ability to rate the acceptability of Vehicle terms across different word classes
T9-Metaphor Continuation-R	1	0.60	Good	Y	ability to recognise coherent continuations of metaphoric discourse
VYesNo	1	0.53	Fair	Y	vocabulary size from 0-10k words known using the Yes/No format
T7-Feelings-R	1	0.50	Fair	Y	ability to recognise metaphors that convey information and feelings about that information
T6-Heuristic-R	1	0.49	Fair	N	ability to recognise similes used to perform heuristic functions
Word Associates Test	1	0.46	Fair	Y	depth of vocabulary knowledge, how well words are known
T2-Metaphor Layering-R	1	0.43	Poor	N	ability to understand layers of figurative and literal meaning in metaphors and puns
T7-Feelings-P	1	0.38	Poor	N	ability to recall metaphors that convey information and feelings about that information
T5-Topic Transition-P	1	0.33	Poor	N	ability to recall proverb/idioms in topic transition in interactive discourse
T6-Heuristic-P	2	0.90	Excellent	N	ability to recall similes to perform heuristic functions
T7-Feelings-P	2	0.31	Poor	N	ability to recall metaphors that convey information and feelings about that information
T8-Idiom Extension-P	3	0.89	Excellent	Y	ability to create possible extensions of idioms
T9-Metaphor Continuation-P	3	0.60	Good	N	ability to create coherent continuations of metaphoric discourse
T8-Idiom Extension-R	3	0.58	Good	N	ability to recognise possible extensions of idioms
T5-Topic Transition-P	3	0.31	Poor	N	ability to recall proverb/idioms in topic transition in interactive discourse
T4-Topic/Vehicle-R	4	0.56	Good	N	ability to rate acceptability of Topic and Vehicle combinations in the framework of an analogy
T5-Topic Transition-R	4	0.54	Fair	N	ability to recognise proverb/idioms in topic transition in interactive discourse

As an NNS+NS factor, the structure of this factor is complicated by the 'good' and 'fair' loadings of Test 9-Metaphor Continuation-R, VYesNo, Test 7 Feelings-R, Test 6-Heuristic-R, and the Word Associates Test. These additional loadings can probably be explained by the fact that the NNS+NS EFA had a smaller number of factors, which tends to result in more variables loading onto each of the factors, or simply that when NSs are involved, **English Grammatical Metaphoric Competence** affects scores for these tests too. The reliability analysis of NNS+NS factors (not presented due to space limitations) showed that VYesNo scores had a huge impact on the internal consistency of **English Grammatical Metaphoric Competence** tests; its inclusion meant extremely low consistency of tests-within-factor (α = 0.172), but its exclusion led to extremely high internal consistency (α = 0.917). This seems to provide further evidence for considering **English Vocabulary Size** as its own construct, even if this factor did not emerge in the NNS+NS EFA. Alpha values for items-within-test were even higher (mean group 1 and group 2 α = 0.960).

The 'misfit' of VYesNo in the NNS+NS factor solution and the fact that it loaded on its own separate factor in the NNS EFA suggest that it should perhaps be removed from this factor, and the EFA subsequently rerun. On the other hand, despite its incongruence with other loading variables, the NNS+NS factor solution revealed VYesNo to be a 'fair' and 'stable' marker of **English Grammatical Metaphoric Competence**; to discard it, therefore, would be to ignore a useful piece of the puzzle. Since the method used to calculate factor scores for the MANOVA (see section 7.4) weighted the relative importance of loading variables, on balance, it was decided that there were more reasons to keep VYesNo as a loading variable for this factor than discard it and rerun the analysis.

NNS+NS factor 2 had the exact same two loading variables as NNS factor 4, and so was identified as **English Illocutionary Metaphor Production**. At the test-within-factor level, tests were moderately positively correlated and significant at the .01 level (r = .50, N = 143); at the items-within-factor level, items displayed a reasonable degree of internal consistency (α = 0.701).

NNS+NS factor 3 matched NNS factor 6, **English Metaphor Language Play**, and was found to have high internal consistency both for tests-within-factor (α = 0.858) and items-within-factor (α = 0.915).

NNS+NS factor 4 matched NNS factor 5, **English Topic/Vehicle Acceptability**, and showed a medium, positive correlation between tests, significant at the .01 level (r = .39, N = 139) and 'moderate' internal consistency of items-within-test (α = 0.560). There are some minor variations between the NNS+NS and NNS solutions for these factors. For instance, in the NNS+NS factor structure Test 3-Vehicle Acceptability-R does not load to **English Topic/Vehicle Acceptability**, and Test 5-Topic Transition-P loads poorly on **English Metaphor Language Play** as an additional variable. The latter finding is perhaps explained by the fact that Test 5-Topic

Transition-P involves thinking up proverbs, phrases or idioms, and thus a certain amount of language play.

NNS+NS factors were all positively correlated. The highest correlation was between **English Grammatical Metaphoric Competence** and **English Metaphor Language Play** (*r* = .74), and the weakest between **English Illocutionary Metaphor Production** and **English Topic/Vehicle Acceptability** (*r* = .34).

7.3.4.3 Calculating factor scores: Dependent variables for MANOVA

In section 7.4 (to follow), a MANOVA is conducted to identify NNS and NS differences on a linear combination of the factors. In order to test for these differences, factors scores were needed. Factor scores are statistical summaries of each participant's performance on each of the factors. Because factors were not measured directly, they need to be estimated from participants' scores on constituent variables and the relative importance of variables-to-factors. Both simple and more sophisticated techniques exist for doing this, all with advantages and disadvantages (Grice, 2001; Revelle, 2017). A 'quick and dirty' method involves standardising scores and then summing those that load highly on each factor. Preferable, however, are more sophisticated techniques such as the Bartlett, Anderson-Rubin, and Regression methods.

With the Bartlett method, scores correlate less well with their own factors but are unbiased and do not correlate with other factors. Factor scores may still correlate with each other. With the Anderson-Rubin method, used with orthogonal rotation only, factor scores are uncorrelated with each other even if factors are correlated (Revelle, 2017). This is the best approach when uncorrelated scores are required. However, since factors were expected and indeed found to be correlated, Thurstone's (1935) Regression method was used to generate factor scores in R. Relative to other sophisticated methods, the regression method yields the highest correlations between factors and factor scores, although chance correlations between variables cause bias so that estimates are sometimes too close to 'true' factor scores (i.e., include less error than they should) (Tabachnick & Fidell, 2013). The regression method has the advantage of maximising validity through factors scores being correlated to the estimated factor. This method generates factor score estimates standardised to a mean of zero, and (since principal axis methods were used) with a standard deviation of factor scores for each factor equal to the squared multiple correlation between factors and variables (DiStefano, Zhu, & Mîndrilă, 2009).

7.4 MANOVA: Exploring L1-L2 group differences on factors

The EFA of the NNS+NS data revealed four underlying and sufficiently distinct competences (i.e.,

factors) that explained 61% of the variance in participants' scores on the observed metaphoric competence and vocabulary tests. The MANOVA and *t*-tests in this section develop this enquiry by identifying the extent of NNS and NS differences in the underlying competences (i.e., scores for each of the factors). In this section I use the terms 'dependent variables' (DVs) to refer to the four NNS+NS factors, and 'independent variable' (IV) to the L1 of participants, which had two categorical levels: Chinese (NNSs) and English (NSs).

7.4.1 Data screening

Although MANOVA is known to be reasonably robust to modest violations of normality and controls and adjusts for type I error (Pallant, 2013; Tabachnick & Fidell, 2013), it nonetheless carries a number of assumptions. Data were therefore screened accordingly using Pallant's (2013) suggested checklist.[81] All criteria pertain to this source unless otherwise indicated.

Data screening showed that:

- Because MANOVA uses 'listwise' deletion of cases, the initial sample size of 143 (NNSs = 112, NS =31) was reduced to 121 for the final analysis (NNSs = 99, NSs = 22), and thus met the criterion of "more cases in each cell than dependent variables" (Pallant, 2013, p. 295);[82]
- Some deletions of participant outliers were necessary (participant 1A was a 1.5 box-length outlier for factor 2, participants 45A, 56A, 14B, 19B, 20B, 21B were 1.5 box-length outliers for factor 4);
- Data were univariate normal evidenced by Kolmogorov-Smirnov and Shapiro-Wilk tests of normality and likely to be multivariate normal evidenced by MVN tool applications (section 7.2.1);
- There were no multivariate outliers evidenced by the fact that the highest Mahalanobis distance of 10.059 (participant 16A) did not exceed the cut-off value of 18.467, the critical value of chi-square when the degree of freedom = 4 ($p < .001$);
- Scatterplot matrixes showed no obvious signs of non-linearity (i.e., horseshoe shapes, S-shapes, curves) and so this assumption was deemed to have been met;
- There was no multicollinearity in the data shown by no overtly high correlations;
- Although Test 5-Topic Transition-P and Test 7-Feelings-P loaded above 0.3 on more than one factor and so contributed to the factor scores of more than one DV, their relative importance to each of their respective factors was small, and the moderate-

[81] Since a 'stepdown analysis' was not performed in the present study, Pallant's (2013) suggestion to check homogeneity of regression did not apply.
[82] An absolute minimum by this criterion is 40 participants in total (2 x levels of the IV x (4DVs + 1)).

to-low correlations between factor scores did not provide evidence of singularity;

- The Box's M Test of Equality of Covariance Matrices is produced during the MANOVA and is reported below

7.4.2 Results

7.4.2.1 MANOVA

NNS and NS group differences on a linear combination of factors 1-4 (DVs)

The mean, standard deviation, minimum and maximum NNS+NS factor scores for NNSs and NSs are shown numerically in Table 7.9. As a feature of the regression method used, factor score estimates are standardised to a mean of zero. Factor 1 **English Grammatical Metaphoric Competence** and Factor 3 **English Metaphor Language Play** displayed the widest distributions of scores, followed by Factor 2 **English Illocutionary Metaphor Production** and Factor 4 **English Topic/Vehicle Acceptability**. A comparison of the NNS and NS factor scores reveals comparatively wider distributions of scores for the NNSs and high NS scores for Factor 1 **English Grammatical Metaphoric Competence**.

Table 7.9 *Descriptive Statistics: NNS+NS Factor Scores, 1-4 (DVs)*

Factor (DV)	*N*	*Min.*	*Max.*	*M*	*SD*
			NNS+NSs		
F1) EGMC	121	-2.81	3.06	-0.13	1.29
F2) EIMP	121	-2.28	2.54	0.07	1.01
F3) EMLP	121	-3	2.93	0.07	1.24
F4) ETVA	121	-1.56	1.7	0.03	0.7
Valid N (listwise)	121				
			NNSs (L1 Chinese) only		
F1) EGMC	99	-2.81	2.32	-0.49	1.12
F2) EIMP	99	-2.28	2.54	0.08	1.08
F3) EMLP	99	-3	2.93	0.06	1.33
F4) ETVA	99	-1.56	1.7	-0.02	0.74
Valid N (listwise)	99				
			NSs (L1 English) only		
F1) EGMC	22	0.81	3.06	1.51	0.55
F2) EIMP	22	-1.73	0.84	0.03	0.67
F3) EMLP	22	-1.84	1.27	0.09	0.77
F4) ETVA	22	-0.62	1.24	0.25	0.4
Valid N (listwise)	22				

Note. Key: EGMC = English Grammatical Metaphoric Competence; EIMP = English Illocutionary Metaphor Production; EMLP = English Metaphor Language Play; ETVA = English Topic/Vehicle Acceptability.

The first stage of interpreting the MANOVA results involved examining the Box's Test of Equality of Covariance Matrices to determine whether the data violated the assumption of homogeneity of variance-covariance matrices (Table 7.10).

Table 7.10 *Box's Test of Equality of Covariance Matrices, NNS+NS Factors 1-4 (DVs)*[a]

Box's M	61.630
F	5.717
*df*1	10
*df*2	6363.873
Sig.[a]	.000

[a] Design: Intercept + L1 group.

[b] Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

The significant result indicates that the assumption was violated and the robustness of the MANOVA not guaranteed. Tabachnick and Fidell (2013, p. 294) suggest that, in such instances, "if cells with larger sample sizes produce larger variances and covariances, the alpha level is conservative, and so the null hypothesis can be rejected with confidence" meaning that the significant finding can be trusted (A Field, 2013). The authors, also suggest using Pillai's criterion instead of Wilke's Lambda for a more robust evaluation of multivariate significance. The variances and covariances for the present data were checked (Table 7.11 and Table 7.12):

Table 7.11 *Variances: NNS and NS Factor Scores, 1-4 (DVs)*

	N	Factor 1	Factor 2	Factor 3	Factor 4
NNSs (L1 Chinese)	99	1.288	1.220	1.753	0.548
NSs (L1 English)	22	0.307	0.453	0.599	0.159

Table 7.12 *Covariances: NNS and NS Factor Scores, 1-4 (DVs)*[a]

	NNSs (L1 Chinese), *n* = 99				NSs (L1 English), *n* = 22			
	1	2	3	4	1	2	3	4
1. Factor 1	1.288	0.354	0.961	0.336	0.307	0.181	0.263	0.002
2. Factor 2		1.220	0.346	0.002		0.453	0.142	0.005
3. Factor 3			1.753	0.006			0.599	0.115
4. Factor 4				0.548				0.159

[a] absolute value (i.e., + or - sign disregarded).

All variances, and all except two covariances (F2-F4, F3-F4), were larger for the larger group, namely the NNSs (L1 Chinese). This provided an adequate indication that the null hypothesis could be rejected, and further analyses trusted.

Next, Levene's Test of Equality of Error Variances was checked (Table 7.13). These results revealed significant values at the .01 level for all four DVs, indicating that this assumption was met.

Table 7.13 Levene's Test of Equality of Error Variances: NNS+NS Factors 1-4 (DVs)[a]

	F	*df*1	*df*2	Sig.
Factor 1	10.118	1	119	.002
Factor 2	6.920	1	119	.010
Factor 3	11.564	1	119	.001
Factor 4	10.290	1	119	.002

Note. Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

[a] Design: Intercept + l1.group.

The multivariate tests of significance showed a statistically significant difference between the NNSs (L1 Chinese) and NSs (L1 English) on the combined dependent variables, $F(4,116) = 111.511$, $p < .001$; Pillai's trace = 0.79, partial eta squared = 0.79. This indicated a difference between the two groups in terms of their metaphoric competence on a linear combination of the four DVs (factors).

Table 7.14 *Multivariate Tests: NNS+NS Factors 1-4 (DVs)*[a]

	Effect	Value	*F*	Hypoth. *df*	Error *df*	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.592	42.068[b]	4.000	116.000	.000	.592
	Wilks' Lambda	.408	42.068[b]	4.000	116.000	.000	.592
	Hotelling's Trace	1.451	42.068[b]	4.000	116.000	.000	.592
	Roy's Largest Root	1.451	42.068[b]	4.000	116.000	.000	.592
L1 group	Pillai's Trace	.793	111.151[b]	4.000	116.000	.000	.793
	Wilks' Lambda	.207	111.151[b]	4.000	116.000	.000	.793
	Hotelling's Trace	3.833	111.151[b]	4.000	116.000	.000	.793
	Roy's Largest Root	3.833	111.151[b]	4.000	116.000	.000	.793

[a] Design: Intercept + L1 group

[b] Exact statistic

NNS and NS group differences on individual factors 1-4 (DVs)

In order to explore whether the NNSs and NSs differed on all the factors, or just some, data from a series of univariate ANOVAs testing between-subjects effects were analysed (Table 7.15). The chance of a type I error (i.e., finding a significant result when there is not really one) was reduced via a Bonferroni adjustment to the alpha level, from .05 to .0125, the original level divided by the number of analyses performed, namely four (Pallant, 2013). Examining the between-subjects effects for L1 group, the only difference to reach statistical significance was factor 1 **English Grammatical Metaphoric Competence**, $F(1,119) = 66.70$, $p < .001$.

Table 7.15 *Test of Between-Subject Effects: NNS+NS Factors 1-4 (DVs)*

Source	Dependent Variable	Type III Sum of Squares	*df*	Mean Square	F	Sig.	Partial Eta Squared

Corrected Model	F1) EGMC	72.21[a]	1	72.206	66.70	.000	.359
	F2) EIMP	.05[b]	1	.054	.05	.820	.000
	F3) EMLP	.02[c]	1	.015	.01	.921	.000
	F4) ETVA	1.32[d]	1	1.321	2.73	.101	.022
Intercept	F1) EGMC	18.90	1	18.901	17.46	.000	.128
	F2) EIMP	.22	1	.219	.21	.647	.002
	F3) EMLP	.43	1	.433	.28	.599	.002
	F4) ETVA	.91	1	.914	1.89	.172	.016
L1 group	F1) EGMC	72.21	1	72.206	66.70	.000	.359
	F2) EIMP	.05	1	.054	.05	.820	.000
	F3) EMLP	.02	1	.015	.01	.921	.000
	F4) ETVA	1.32	1	1.321	2.73	.101	.022
Error	F1) EGMC	128.83	119	1.083			
	F2) EIMP	123.53	119	1.038			
	F3) EMLP	185.22	119	1.556			
	F4) ETVA	57.52	119	.483			
Total	F1) EGMC	202.93	121				
	F2) EIMP	124.22	121				
	F3) EMLP	185.80	121				
	F4) ETVA	58.93	121				
Corrected Total	F1) EGMC	201.04	120				
	F2) EIMP	123.59	120				
	F3) EMLP	185.24	120				
	F4) ETVA	58.84	120				

Note. Key: EGMC = English Grammatical Metaphoric Competence; EIMP = English Illocutionary Metaphor Production; EMLP = English Metaphor Language Play; ETVA = English Topic/Vehicle Acceptability.

[a] R^2 = .359 (Adjusted R^2 = .354).

[b] R^2 = .000 (Adjusted R^2 = -.008).

[c] R^2 = .000 (Adjusted R^2 = -.008).

[d] R^2 = .022 (Adjusted R^2 = .014).

The effect that L1 group had on **English Grammatical Metaphoric Competence** is shown via the effect size statistics in the far right column. SPSS reports partial eta squared, an effect size statistic measuring the proportion of total variance in each DV (factor) associated with membership of the IV group (i.e., L1 Chinese or L1 English), with the effects of any other IVs and interactions partialled out (Richardson, 2011). As a one-way design with only one IV, here the partial eta squared statistics are identical to eta squared statistics (Norouzian & Plonsky, 2018).[83] The value = 0.36, namely 36% of the variance in **English Grammatical Metaphoric Competence**, a very 'large' amount indeed (J. Cohen, 1988; Tabachnick & Fidell, 2013).

[83] Regardless, SPSS (v. 24) labelled this statistic 'partial eta squared'. Confusion between these two terms is a persistent problem in the SLA field, trace-able in part to the fact that earlier versions of SPSS (and related handbooks) mislabelled eta squared and partial eta squared (Loewen et al., 2014).

7.4.2.2 Independent-samples t-tests: NNS+NS factors 1-4 (DVs)

A second approach to examining differences between the L1 groups on the DVs (factors) was to use independent-samples *t*-tests, Cohen's *d* measure of effect size, and 95% confidence intervals. The assumptions of scale measurement, random sampling from the population, independence of observations, and normal distribution were met in the data and so independent-samples *t*-test was chosen over the non-parametric equivalent. Table 7.16 presents the number of participants included in each group,[84] group means, standard deviations, and standard errors of the mean.

Table 7.16 *Group Statistics (Independent Samples Test): NNS+NS Factors 1-4 (DVs)*

Factors 1-4 (DVs)	L1 group	*N*	*M*	*SD*	*SE* (of *M*)
F1) English Grammatical Metaphoric Competence	L1 Chinese	106	-.4344	1.1347	.1102
	L1 English	22	1.5138	.5540	.1181
F2) English Illocutionary Metaphor Production	L1 Chinese	105	.0218	1.1047	.1078
	L1 English	22	.0278	.6731	.1435
F3) English Metaphor Language Play	L1 Chinese	106	.0441	1.3240	.1286
	L1 English	22	.0921	.7737	.1650
F4) English Topic/Vehicle Acceptability	L1 Chinese	100	-.0255	.7402	.0740
	L1 English	22	.2481	.3993	.0851

The significant results for Levene's Test for Equality of Variances indicated that data in the *t*-test should interpreted from the 'equal variances not assumed' rows for all four factors. A very "large" effect (Plonsky & Oswald, 2014, p. 889) for L1 group difference, significant at the .01 level, was found for Factor 1 **English Grammatical Metaphoric Competence**, $t(63.81) = -12.059$, $p < .001$ (two tailed), Cohen's $d = -2.18$, 95% confidence interval (-2.72, -1.64). An apparently "small" effect (Plonsky & Oswald, 2014, p. 889), significant at the .05 level, was found for Factor 4 **English Topic/Vehicle Acceptability**, $t(57.77) = -2.426$, $p = .018$ (two tailed), Cohen's $d = -0.46$, 95% confidence interval (-0.93, 0.01). However, since the 95% confidence intervals pass through zero, this effect is likely to be negligible. No effect or statistically significant L1 group difference

[84] Independent samples *t*-test allows for cases to be excluded 'pairwise' (analysis-by-analysis), meaning the number of participants was slightly higher than in the MANOVA, which automatically implements 'listwise' deletion of cases.

Table 7.17 *Independent Samples Test: NNS (L1 Chinese) and NS (L1 English) Group Differences*

		Levene's Test for Equality of Variances		*t*-Test for Equality of Means							Effect size		
									95% CI of Mean Difference			95% CI of effect size[a]	
		F	Sig.	*t*	*df*	Sig. (2-tailed)	Mean Difference	*SE* Difference	Lower	Upper	Cohen's *d*[b]	Lower	Upper
F1) English Grammatical Metaphoric Competence	Equal variances assumed	10.132	.002	-7.843	126	.000	-1.94823	.24840	-2.43979	-1.45666			
	Equal variances not assumed			-12.059	63.808	.000	-1.94823	.16155	-2.27098	-1.62547	-2.181	-2.720	-1.640
F2) English Illocutionary Metaphor Production	Equal variances assumed	7.297	.008	-.025	125	.980	-.00600	.24495	-.49079	.47879			
	Equal variances not assumed			-.033	48.286	.973	-.00600	.17948	-.36682	.35482	-0.007	-0.470	0.460
F3) English Metaphor Language Play	Equal variances assumed	11.188	.001	-.164	126	.870	-.04803	.29268	-.62724	.53118			
	Equal variances not assumed			-.230	50.554	.819	-.04803	.20916	-.46802	.37196	-0.044	-0.510	0.420
F4) English Topic/Vehicle Acceptability	Equal variances assumed	10.100	.002	-1.677	120	.096	-.27362	.16314	-.59663	.04940			
	Equal variances not assumed			-2.426	57.767	.018	-.27362	.11281	-.49944	-.04779	-0.459	-0.930	0.010

[a] Calculated using syntax developed by Jeromy Anglim (2016) for R package 'compute.es' (Del Re, 2013).
[b] Calculated using Becker's (2000) Effect Size Calculators available at: http://www.uccs.edu/~lbecker/

was found for factor 2 **English Illocutionary Metaphor Production**, $t(48.29) = -0.033$, $p = .973$, Cohen's $d = -0.01$, 95% confidence interval (-0.47, 0.46) or factor 3 **English Metaphor Language Play**, $t(50.55) = -0.230$, $p = .819$, Cohen's $d = -0.04$, 95% confidence interval (-0.51, 0.42).

Summarising these findings, the MANOVA showed that the NNS (L1 Chinese) and NSs (L1 English) displayed differences on a linear combination of the DVs (factors). A series of ANOVAs showed statistically significant differences between groups on the first factor only, namely **English Grammatical Metaphoric Competence**. The independent-samples *t*-test also revealed that NNSs and NSs were most distinguishable by their **English Grammatical Metaphoric Competence**. The small effect observed for **English Topic/Vehicle Acceptability** was considered negligible, since 95% confidence intervals passed through zero. These findings are discussed in detail in the next chapter.

7.5 English Grammatical Metaphoric Competence: How does form frequency relate to item difficulty and discriminability? A case study of phrasal verbs

To probe the NNS and NS group difference for **English grammatical metaphoric competence**, the relationships between frequency of form and item difficulty and discriminability in Test 1-Phrasal Verbs-R and –P (the strongest marker variables for **English grammatical metaphoric competence**) were investigated. Frequency of form refers to how often phrasal verbs, with all their various senses, appear in language.

The first step in this analysis was to determine the form frequency of each of the 20 phrasal verb test item. Table 7.18 shows each verb indexed with BNC-BYU frequencies and BNC lemma frequencies (reported in D. Gardner & Davies, 2007). Next, correlations between these frequencies and item difficulty and discriminability indexes for receptive and productive tests were calculated. Table 7.19 shows these data. Although a 'medium' positive correlation between BNC lemma frequencies and receptive item difficulty was found ($r = .30$), this was non-significant at the .05 level and negligible, since lower and upper 95% confidence intervals passed through zero. Although confidence intervals did not pass through zero for correlations between form frequency (both indexes) and Test 1-Phrasal Verbs-R item discriminability, the fact that the lower bound is close to zero (i.e., indicating no relationship) and correlations were not significant at the .05 level suggests that the relationship may have been negligible.

Taken together, these results show that in both the receptive and productive modes, there was no association between the frequency of the 20 phrasal verb forms and their ease or difficulty, nor between form frequency and the ability of phrasal verbs to discriminate between higher and lower ability test takers.

Table 7.18 *Frequency of 20 Phrasal Verb Forms*

No.	Item	Frequency form BNC-BYU	BNC lemma freq. Gardner & Davies, 2007)
1	pick up	7391	9037
2	take down	312	775
3	get out	4959	3545
4	give out	472	532
5	put off	543	743
6	get in	4105	1127
7	put down	1364	2873
8	take off	2221	2163
9	break up	1302	1286
10	come in	9925	4814
11	go down	5250	4781
12	pick out	811	856
13	move up	594	477
14	get down	935	1538
15	put in	3705	810
16	hold up	1397	1624
17	get off	1426	1086
18	bring in	2353	2505
19	go out	8496	7688
20	come off	1188	518

Table 7.19 *Correlations: Form Frequency, Item Difficulty and Discriminability (20 Phrasal Verbs)*

	Test 1-Phrasal Verbs-R		Test 1-Phrasal Verbs-P	
	Item diff. (p)	Item discr. (D)	Item diff. (p)	Item discr. (D)
BNC-BYU frequency	.03 (-.41,.60)	-.44 (-.78,-.05)	.00 (-.50,.60)	-.24 (-.56,.17)
BNC Lemma (D. Gardner & Davies, 2007)	.30 (-.16,.68)	-.43 (-.75,-.01)	.24 (-.37,.73)	-.23 (-.57,.26)

Note. No correlations significant at the .01 or .05 levels. Lower and upper 95% confidence intervals from 1,000 bootstraps reported in brackets. Correlation between BNC-BYU and BNC lemma frequency = .84 (.70,.96), $p < .01$.

The validity of the two frequency indexes, the BNC-BYU and BNC lemma (from D. Gardner & Davies, 2007) is evidenced by their 'large' correlation, significant at the .01 level (r = .84), and confidence intervals not passing through zero, which show that the indexes corroborate one another.

7.6 Chapter summary

The focus of this chapter was on investigating:

a) the extent to which factors underlie the observed L2 MC, vocabulary knowledge and proficiency test scores for NNSs and the kind of (sub)competences that these factors represented (RQ3)
b) the extent to which the same factors can be found in the NNS and combined NNS+NS data, and how NNSs (L1 Chinese) and NSs (L1 English) factor scores differ (RQ4)

The EFA of the NNS data showed that 42% of the total variance in MC test, vocabulary test and

proficiency test components (23 variables) was explained by a statistically adequate six-factor solution. Using information about loading strength and stability across 5,000 bootstrap resamples, and descriptions of the skills tested, factors were interpreted as **(F1) English Vocabulary Size**, **(F2) English General Comprehension**, **(F3) English Grammatical Metaphoric Competence**, **(F4) English Illocutionary Metaphor Production**, **(F5) English Topic/Vehicle Acceptability**, and **(F6) English Metaphor Language Play.**

The EFA of the NNS+NS data showed that a statistically adequate four-factor solution explained 61% of the Total Variance in NNSs and NSs combined MC and vocabulary test scores (17 variables). The four factors corresponded with four out of six found in the NNS data: (NNS+NS F1, NNS F3) **English Grammatical Metaphoric Competence**, (NNS+NS F2, NNS F4) **English Illocutionary Metaphor Production**, (NNS+NS F3, NNS F6) **English Metaphor Language Play,** and (NNS+NS F4, NNS F5) **English Topic/Vehicle Acceptability**. Since fewer variables had been submitted for analysis, and the NS data also included, differences in loading variables were observed. Factor scores were then calculated for these four factors using Thurstone's regression method.

A MANOVA showed that NNSs (L1 Chinese) and NSs (L1 English) statistically differed on a linear combination of the four factors. Concerning NNS and NS differences on the four factors individually. a series of univariate ANOVAs showed and independent-samples *t*-test showed that groups differed on factor 1 **English Grammatical Metaphoric Competence** only. There were not statistical differences for the other three factors.

Finally, to probe NNS and NS group differences for factor 1 **English Grammatical Metaphoric Competence** further, the relationships between frequency of form and item difficulty and discriminability in Test 1-Phrasal Verbs-R and –P (the strongest marker variables for English grammatical metaphoric competence) were investigated. The correlation analysis showed no statistical relationships between frequency of phrasal verb forms and item difficulty and discriminability scores for either receptive or productive tests.

Chapter 8: Discussion of Analysis 2

8.1 Introduction

This chapter will discuss the key findings of Analysis 2: Metaphoric and other (sub)competences uncovered in relation to previous metaphoric competence and other research. The discussion is again structured into two main parts corresponding to the third and fourth research questions, with subsections on emerging themes.

8.2 RQ3: To what extent do factors underlie the observed L2 metaphoric competence, vocabulary knowledge and proficiency test scores for NNSs? What kind of (sub)competences might these factors represent?

8.2.1 The process of discovering L2 metaphoric (sub)competences

In order to explore and uncover underlying factors (RQ3), Exploratory Factor Analysis was used. Because the aim was *not* to reduce or consolidate variables or measure causal relationships of empirically and theoretically established MC (sub) competences, Principal Components Analysis (PCA), Confirmatory Factor Analysis (CFA) and Structural Equation Modelling (SEM) were unsuitable. These distinctions may seem trivial, but it is important to state them clearly, first in order to properly understand how the present study's results relate to those of past research, and second so as not to over- or misinterpret findings.

The NNS EFA produced a six-factor solution, explaining 42% of variance in 23 MC, vocabulary and general proficiency tests. The resulting model had 'fairly' consistent factors, and was adequate by all post-hoc criteria (Appendix H). The marker (highest loading) variables for five out of six factors were 'stable' and highly likely to load on the same variables in replication of the analyses, evidenced by their 95% confidence interval lower bounds. The six factors provided empirical support for the existence of several (sub)competences affecting scores on the MC, vocabulary knowledge, and general proficiency tests. At this stage the model is exploratory, however, the bootstrap loadings suggest that the strongest marker variables would very likely replicate on a comparable sample of NNSs. To the best of my knowledge, this is the first time that EFA has been used as a technique in L2 metaphoric competence research.

The present study took several steps to improve basic methodological issues in past studies of L1 metaphoric competence (particularly H. R. Pollio & Smith, 1980). These included using a factor analysis method appropriate to the research question (Plonsky & Gonulal, 2015); reporting measures of sample adequacy and use of a robust method for increasing power;

reporting relevant information about participants' backgrounds; reporting measures of model adequacy; providing estimates on the reliability of instruments, rater decisions and factors yielded; basing qualitative descriptors of loadings (e.g., 'good', 'high', strong', 'stable') in empirical resamples data, and using descriptors consistently throughout the report (cf. H. R. Pollio & Smith, 1980). Attention to these issues strengthened the methodology and substantive findings of the present study, and hopefully sets a precedent for future investigations into L2 MC.

In this section, two main issues are discussed: 1) how the 'numbers' (e.g., factor-to-variable ratio, total variance explained) of the NNS EFA compared with those of past research, and how this informs our understanding of L2 metaphoric competence; 2) the extent to which the approach to factor retention shaped the results.

8.2.1.1 Present and past research: Comparing the numbers

The NNS EFA yielded six factors from 23 variables, a factor-to-variable proportion of 0.26. While there is unfortunately no equivalent L2 EFA to compare this value with, it lies roughly halfway between Pollio and Smith's (1980) factor-to-variable proportion of 0.15 from a PCA on (adult) L1 metaphoric competence and complex human problem solving test scores, and Mashal and Kasirer's (2012) value of 0.36 from a PCA on L1 comprehension of visual and verbal metaphor test scores in differently developed children. The proportion also matches the value of 0.26 observed by Beaty and Silvia (2013) and Silvia and Beaty (2012), who used structural equation modelling to investigate L1 metaphoric competence and general intelligence. Although the matter is complicated by the fact that the Beaty and Silvia's proportion comes from a CFA and the other authors misused PCA (rather than EFA) to underlying traits in the data, the present study's finding and those of past research provisionally suggest that if one is factor analysing between 20 and 30 (L1 or L2) metaphoric competence variables, a four-to-six factor solution is likely to be the most parsimonious and interpretable.

Although the Principal Component solutions in these past studies explained more total variance than the NNS EFA (in some cases > 80%), there are two things to support the theoretical usefulness of the 42% of total variance explained by the six-factor solution. First, PCA inherently explains more variance than EFA,[85] since the former does not differentiate between variance that is shared versus unique among variables, but the latter does. Second, based on the findings of meta-analytic work, Plonsky and Gonulal (2015) have suggested a revision (i.e., potential lowering) of Field's (2013) 55-65% benchmark, particularly when certain 'difficult-to measure' constructs are under investigation.

[85] If PCA had been used in the present study, the six-component solution would explain 57% of the total variance.

8.2.1.2 To what extent did the approach to factor retention shape the results?

Methodological decisions on which tests to submit for analysis, the factor extraction and rotation method, and retention criteria used all have a large impact on eventual EFA solutions. Factor retention is a particularly problematic issue in this respect, evidenced in both past and present research. Pollio and Smith (1980) for instance, tried out different numbers of components before forcing a five-component solution, eventually reduced to four, since one component required re-classification as part of another. In the few examples of PCA in L1 MC research found, more robust methods such as parallel analysis and bootstrapping of eigenvalues were not explored and considered to help determine the extent to which the data can be reduced to core components, or which underlying factors are present. However, the number of factors to retain is less of a problem in SEM, primarily because one begins the analysis with a pre-existing theory and model. Although CFA (conducted as part of SEM) uses comparatively more robust methods for detecting patterns of relationship in variables, it is not immune to the problem of detecting superfluous constructs (Patil, Singh, Mishra, & Donavan, 2008).

In the present study, the strictest factor retention criterion, parallel analysis (Horn, 1965), suggested the retention[86] of one or two factors, whereas the most liberal, Joliffe's (1972) > 0.7 rule, suggested retaining 12. This discrepancy, in itself, suggests the need for further research into the suitability of different factor retention criteria in L2 MC research, and variables that may influence estimates. A one-factor solution has the advantage of a high degree of parsimony but a low amount of total variance explained; in the 12-factor solution, the advantages and disadvantages are reversed. Since a 12-factor solution results in several variables with one marker only and is hardly interpretable, it is not worth considering. A one-factor solution, one the other hand, presents the possibility of a highly parsimonious model of L1 MC, vocabulary knowledge and general proficiency, which could be assumed to be overall L2 general proficiency or overall L2 MC. So why interpret the less parsimonious six-factor solution?

Although a one-factor solution was intriguing, it was too reductive for exploring potential (sub)competences of L2 MC. Moreover, the validity of the six-factor solution is bolstered by its interpretability, the loading of receptive and productive versions of some tests (e.g., Phrasal verbs and Idiom extension) on the same factors, and the fact that two of the factors (**English Vocabulary Size** and **English General Comprehension**) were largely conceptually independent from MC[87] and most strongly correlated with one another, rather than MC factors.

[86] A one-factor solution (conducted in SPSS) on the same 23 variables and using the extraction and rotation explains 25% total variance after extraction, a 12-factor solution explains 57%.

[87] In the NNS+NS EFA, the VYesNo loaded as a 'fair' and 'stable' marker of **English Grammatical Metaphoric Competence**. However, its conceptual distinctness from this (and other MC) factors is evidenced by its negative impact on the internal consistency of tests for this factor: $\alpha = .172$ with the VYesNo, and $\alpha = .917$ without it.

Given that the respective strongest marker variables for these factors, VYesNo and IELTS Listening, are established measures of non-MC constructs, this latter finding is somewhat reassuring. The six-factor solution, therefore, allowed for two basic conclusions: 1) L2 MC, **English Vocabulary Size** and **English General Comprehension** appear to be distinct constructs; 2) L2 MC itself has several (sub)competences. Concerning the first finding, the fact that vocabulary size emerged as a strong marker of the 'vocabulary' factor (and the WAT did not load at all) aligns with past research showing vocabulary size to be a purer marker of vocabulary knowledge than depth, which presumably has more conceptual overlap with other areas of language competence (Tseng & Schmitt, 2008; D. Zhang, 2012). The second basic conclusion is discussed below.

8.2.2 The nature of L2 metaphoric (sub)competences

The other part of research question three sought to understand the nature of underlying (sub)competences (i.e., factors) in relation to their loading variables (i.e., tests). Since the conceptual distinctiveness of **English Vocabulary Size** and **English General Comprehension** has been teased out already, the discussion here focuses on the L2 metaphoric (sub)competences identified.

EFA can be understood in terms of underlying factors causing or affecting scores on observed variables. After considering the strength, population stability and task requirements associated with each loading variable, names were given to factors. These were, with strongest marker variables in parenthesis: **(F1) English Vocabulary Size** (VYesNo, 'excellent' marker); **(F2) English General Comprehension** (IELTS Listening, 'good' marker); **(F3) English Grammatical Metaphoric Competence**, (Test 1-Phrasal Verbs-P, 'good' marker); **(F4) English Illocutionary Metaphor Production** (Test 6-Heuristic-P, 'excellent' marker); **(F5) English Topic/Vehicle Acceptability** (Test 4-Topic/Vehicle-R, 'good' marker); and **(F6) English Metaphor Language Play** (Test 8-Idiom Extension-P 'excellent' marker).

As noted, no known EFA of L2 MC variables (on their own or in combination with non-MC variables) exists. Moreover, comparing the factor structure of the NNS EFA with those of found in L1 MC research is problematised by differences in participant, instrument and study features across various publications. Nevertheless, since metaphor is not a purely linguistic phenomenon (section 2.1), consideration of how the NNS EFA aligns with factor structures uncovered in past L1 metaphoric competence research is justified, particularly with regard to the 'non-linguistic' traits uncovered.

8.2.2.1 Conventional and creative aspects of L2 metaphoric competence

In the literature review, a synthesis of PCA, EFA, CFA and SEM approaches to L1 metaphoric

competence in research on intelligences (Beaty & Silvia, 2013; Silvia & Beaty, 2012) and complex problem solving (H. R. Pollio, 1977; H. R. Pollio & Smith, 1980) led to several speculations. The first was that (sub)competences revealed from a factor analysis of L2 data *might* reflect a conceptual distinction between 'creative' and 'conventional' use of metaphor, since such factors had emerged in both lines of research cited above.

Comparing past and present research in this respect was complicated by two problems. First, although in the present study 'conventionality' and 'creativity' (and the related concept of 'novelty') could be operationalised in terms of dictionary codification (or lack of), none of the studies cited above had taken this approach. Beaty and Silvia (2013) for instance, measured 'conventional' metaphor via a Vehicle term generation task scored by two raters using a six-point aptness scale, and 'creative' metaphor via a simile completion task scored (again) by two raters using scales of Vehicle-Topic remoteness, novelty, and cleverness. Second, what exactly should be counted as 'conventional' or 'creative' metaphor, and how would this relate to 'novelty'? For instance, one test item may elicit production of a 'novel' lexical item for representing a 'conventional' cross-domain mapping (e.g., LOVE IS A JOURNEY), another may measure sensitivity to the acceptability of a 'conventional' metaphor using an 'unconventional' syntactic structure. Both instances contain 'conventional', 'creative', and 'novel' dimensions.

The approach eventually taken for characterising factors in terms of 'creative' and 'conventional' use of metaphor involved consideration of the extent to which the form and syntax of metaphors (i.e., test items or elicited NNS productions) were codified in dictionaries or retrievable in language corpora.

The factor most related to 'creativity' in the present study seemed to be **(F6) English Metaphor Language Play**, defined by the 'excellent' marker Test 8-Idiom Extension-P, which involved producing appropriate and funny extensions of the literal senses of idioms, and also 'poorly' by Test 9-Metaphor Continuation-P and Test 8-Idiom Extension-R. To a large extent, this factor involved the capacity to do something 'novel' with something 'conventional' and concerned the adaption of existing knowledge structures to solve new problems that lack pre-existing, tailor-made linguistic solutions (Candlin, 1986). In this vain, it is similar to Pollio and Smith's (1980) 'innovative figurative use' factor, which was marked by the ability to produce original and figurative noun adjective word associations, simile endings and metaphor symbols. It could also be argued that **(F4) English Illocutionary Metaphor Production** involved 'creative' metaphor, since its markers (both productive tests) did not require specific 'conventional' forms, and since its strongest marker elicited metaphor productions to perform ad-hoc heuristic functions.

The factor most related to conventional metaphor in the present study appears to be **(F2) English Grammatical Metaphoric Competence**. The 'conventionality' of the metaphors

engaged by this factor is evidenced by its loading variables, which elicited knowledge of fixed, grammatical structures and general proficiency. The factor was defined by Test 1-Phrasal Verbs-P ('good' marker), which elicited the production of closed class (metaphorical) particles to form phrasal verbs and offered test takers no 'creative' freedom. According to Beaty and Silvia (2013), knowledge of 'conventional metaphor is rooted in crystallised intelligence, which involves (among other things) vocabulary knowledge, general knowledge and personality. To the extent that it involves 'conventional' metaphor, one would therefore expect these aspects to be good predictors of **English Grammatical Metaphoric Competence.**

The remaining MC factor, **(F5) English Topic/Vehicle Acceptability** seems to involve both a strong 'creative' and 'conventional' component. Because its loading variables all measured the ability to recognise acceptable Vehicles and reject unacceptable ones, with Test 4-Topic/Vehicle-R ('good' marker) and Test 3-Vehicle Acceptability-R ('fair' marker) using empirically established NS norms, to some extent, the factor concerned knowledge of 'conventionality'. However, given that Test 4-Topic/Vehicle-R items were designed around concepts rather than from corpora, it could also be argued that they are 'novel', and thus, that some 'creative' aspects of metaphor were being engaged.

The basic conclusion here is that 'creativity' and 'conventionality' seemed to characterise L2 MC factors in the present study, just as they have been shown to characterise L1 MC factors in previous research. The extent to which each factor can be located on these dimensions is at this stage, speculative. Beaty and Silvia (2013; 2012) showed that for normally developed adult L1ers, the quality of creative metaphors is best predicted by higher-order executive processes, whereas the ability to generate conventional metaphors is best predicted by crystallised knowledge. Although Littlemore (2001) operationalised L2 MC in terms of these higher order processes (fluid intelligence) and explored its relationship to cognitive style, she did not explore how this type of L2 MC predicted metaphor creativity. Future research might therefore seek to understand whether Beaty and Silvia's (2013; 2012) findings on the ability of certain intelligences to predict conventional and creative metaphor use extends to L2 MC.

8.2.2.2 Revisiting Low (1988) and Littlemore and Low (2006a, 2006b)

Low's (1988) ten-skill framework and Littlemore and Low's (2006a) four-component framework convey a superficial sense of order, hierarchy, and conceptual relatedness and independence. But did MC tests, designed to measures these skills and (sub)competences really support the frameworks suggested by the authors?

Factors ranged from homogenous to heterogeneous concerning the extent to which loading variables belonged to the same components of the authors' frameworks. At the more homogenous end of the spectrum, **(F3) English Grammatical Metaphoric Competence**

contained MC tests designed to measure metaphor and Grammatical Competence only, and **(F4) English Illocutionary Metaphor Production** exclusively had tests of metaphor and illocutionary competence (Littlemore & Low, 2006a). However, the fact that **(F3) English Grammatical Metaphoric Competence** also had OOPT variables, and that Test 6-Heuristic-R and Test 7-Feelings-R (receptive measures of metaphor and illocutionary competence) hardly loaded on any factors, meant that the authors' framework components were only dimly reflected in these factors. Moreover, **(F5) English Topic/Vehicle Acceptability, (F6) English Metaphor Language Play** and **(F2) English General Comprehension** (discussed here since it had MC tests as loading variables)[88] were all heterogeneous with respect to loading variables and the authors' frameworks. For instance, the makeup of **(F5) English Topic/Vehicle Acceptability** suggests that it may now be warranted to look for connections between knowledge of the boundaries of conventional metaphor (Low, 1988), of acceptable Topic and Vehicle combinations (Low, 1988) and figurative language in topic transition (Littlemore & Low, 2006a).

Another observable pattern in the factors concerns the modalities they reflect. **(F5) English Topic/Vehicle Acceptability** and **(F2) English General Comprehension** (predominantly, but IELTS Speaking excepted) are marked exclusively by receptive tests. Similarly, **(F4) English Illocutionary Metaphor Production** is marked by productive tests only. The fact that both its constituents, Test 6-Heuristic-P and Test 7-Feelings-P, used items in the form of direct metaphor (i.e., similes), may also point to a syntactic thread within this construct (Glucksberg & Haught, 2006; Haught, 2013). On the other hand, **(F3) English Grammatical Metaphoric Competence** and **(F6) English Metaphor Language Play** involved both receptive and productive knowledge. These findings may suggest that the possession of **(F3) English Grammatical Metaphoric Competence** and **(F6) English Metaphor Language Play** ability tends to equip learners to both understand and produce metaphor in these domains, whereas **(F4) English Illocutionary Metaphor Production** (as the name suggests) **and (F5) English Topic/Vehicle Acceptability** concern skills that are not necessarily transferable across receptive and productive modalities.

In summary, these points suggest that the skill/competence frameworks suggested by Low (1988) and Littlemore and Low (2006a) need to be further scrutinised and, if necessary, revised. Since these publications have made longstanding contributions to research into L2 metaphoric competence, a reconsideration of basic tenets must proceed with caution. Nevertheless, the findings of the present study lead to the conclusion that L2 metaphoric competence (as measured) is underpinned by grammatical, productive illocutionary, Topic/Vehicle acceptability and creative/ludic dimensions, and L2 general comprehension and vocabulary size constitute distinctly separate constructs. These findings do not come close to

[88] **(F1) English Vocabulary Size** is not relevant to this part of the discussion since none of its loading variables were MC tests.

offering a new theoretical model of L2 metaphoric competence, but rather, form a series of informal hypotheses that need to be explored in further research. Although the use of the bootstrapping technique suggested which factors may replicate and in what form, the real proof of the theoretical pudding (or lack of) will come via 'external', rather than 'internal' replication.

8.3 RQ4: To what extent can the same factors be found in the NNS and combined NNS+NS data, and how do the NNSs' and NSs' factor scores differ?

8.3.1 NNS and NNS+NS factors

The EFA of the NNS+NS data had two purposes. First, it allowed for investigation into the extent to which the same factors can be found in the NNS and combined NNS+NS data. This step enabled factor scores to be estimated for all NNS and NS participants. Since the NNS+NS data were treated as a population for this EFA, factor scores for both NNSs and NSs were directly comparable, something which would not have been possible had separate EFAs of the NNS and NS data been conducted. A subsequent MANOVA and independent-samples *t*-test were then used to investigate the extent to which NNSs and NSs differed on both overall and individual factors.

Concerning the first part of research question four, the NNS+NS EFA produced a four-factor solution explaining 61% of the total variance in 17 tests (MC and vocabulary only), which were interpreted as representing, with some differences, the four MC factors found in the NNS EFA: **English Grammatical Metaphoric Competence**, **English Illocutionary Metaphor Production**, **English Metaphor Language Play**, and **English Topic/Vehicle Acceptability**. The model was also adequate by all post-hoc criteria and had even higher test-within-factor and item-within-factor reliability or correlation estimates than for NNS EFA factors. For this solution, the marker variables for two out of four factors were 'stable'.

Concerning the second part of research question four, the MANOVA showed that NNSs (L1 Chinese) and NSs (L1 English) varied on a linear combination of the four factors (i.e., dependent variables), thus indicating that their English metaphoric competence was statistically different. A series of univariate ANOVAs testing effects between NNS and NS groups showed a very 'large', statistically significant difference (at the .01 level) for **English Grammatical Metaphoric Competence** only. This finding was corroborated by a comparison of means, effect sizes and confidence intervals, which again showed NNS (L1 Chinese) and NS (L1 English) group differences for this factor only. Although an apparently 'small', statistically significant (at the .05 level) was found for **English Topic/Vehicle Acceptability**, the bi-polarity of the confidence intervals suggested that this effect was negligible. In other words, NNSs (L1 Chinese) and NSs (L1 English) differed in their scores for **English Grammatical Metaphoric Competence**, but not for

English Illocutionary Metaphor Production, English Metaphor Language Play, or **English Topic/Vehicle Acceptability.**

Since the MANOVA and independent-samples *t*-test results are based on factors uncovered in the NNS+NS EFA, the following discussion focuses on NNS (L1 Chinese) and NS (L1 English) similarities and differences for factors as defined in *this* EFA, rather than one on the NNS data.

8.3.2 L2 English Grammatical Metaphoric Competence: The hardest aspect of L2 metaphoric competence to acquire?

The NNSs (L1 Chinese) and NSs (L1 English) differed statistically in terms of their **English Grammatical Metaphoric Competence** only. Compared with other factors, this one had the lowest scores for the NNSs (L1 Chinese), but the highest scores for the NSs (L1 English). Taken together, these findings suggest that **English Grammatical Metaphoric Competence** may be the hardest aspect of L2 metaphoric competence to acquire. But why?

The three strongest ('excellent') marker variables for **English Grammatical Metaphoric Competence**, Test 1-Phrasal Verbs-R, Test 1-Phrasal Verbs-P, and Test 3-Vehicle Acceptability-R, had one thing in common - phrasal verbs. Although knowledge of phrasal verbs was not the explicit focus of Test 3-Vehicle Acceptability-R, five out of 18 items retained for this test in the NNS+NS data file contained phrasal or prepositional verbs (items 4, 5, 6, 8, 12). The findings of the present study, in combination with those of past research showing patterns of avoidance of phrasal and prepositional verbs at lower L2 proficiency levels and continued problems with figurative phrasal verbs at higher proficiency levels (Liao & Fukuya, 2004), affirm that this area of L2 MC is "a traditional and recurring nightmare for all learners of English" (Littlemore & Low, 2006a, p. 158).

Although it is tempting to conclude that the NNS participants in the present study experienced difficulty with phrasal verbs because their L1 (Chinese) does not contain this aspect of language, the fact that avoidance of English phrasal verbs has been observed in the interlanguage of lower proficiency L2ers from both typologically similar *and* different languages to English, warns otherwise. While NNSs from typologically similar languages may benefit from a degree of positive transfer, other problems may arise. For instance, Dutch learners of English have been found to avoid using (acceptable) figurative English phrasal verbs, because they were perceived to be too Dutch-like (Kellerman, 1983; Liao & Fukuya, 2004), a problem which L2 learners from languages such as Chinese effectively bypass. The implication is that further research involving MC tests and NNSs from different L1 groups is needed to determine the extent to which test items involving phrasal verbs pose universal difficulties for L2ers, and the specific role(s) that typological distance plays.

8.3.2.1 The role (or non-role) of form frequency

Because the MANOVA and independent-samples *t*-tests showed that NNS and NS groups statistically differed in their **English grammatical metaphoric competence** only, a case study of the potential relationships between frequency of form and item difficulty and discriminability in Test 1-Phrasal Verbs-R and –P (the strongest marker variables for **English grammatical metaphoric competence**) was conducted. The correlation analysis (section 7.5) showed no significant relationship (at the .05 or .01 level) between the frequency of phrasal verb forms, and item difficulty or discriminability, either in the receptive or productive mode. This finding speaks to the general view that frequency alone cannot account for some aspects of L2 learning (cf. Ellis, 2002; Gass & Mackey, 2002), and suggests that **English Grammatical Metaphoric Competence** (as measured) is likely to be more strongly affected by numerous other factors (e.g., perceived saliency of forms, representation in the L1, degree of metaphoricity).

Although the validity of these results is bolstered by the use of two independent measures of form frequency, search of the BNC-BYU and BNC lemmas presented in D. Gardner and Davies (2007), their findings are limited in several ways. First, the frequencies counted reflect the 'form' of the phrasal verbs only, not the number of corpus hits with the particular metaphoric 'form-meaning' mapping of the test item, which is likely to be much lower in frequency. Second, although for most phrasal verbs, what might be called the 'infinitive' form is the most frequent (e.g., 'give up' as opposed to 'giving up'), for others, a 'non-infinitive' form is the most frequent (e.g., 'moves up' is more frequent than 'move up'). The point is that English phrasal verbs are not all equal with regard to the frequency and proportions of their various derivations (–s, -ed, -ing, etc.) and this should be assumed to affect their learnability, particularly given that L1 is known affect morpheme acquisition order (Luk & Shirai, 2009).

Understanding the possible role of frequency and proportions of phrasal verb derivations with regard to the test items requires further analysis. Although past research would suggest that this did play a role (Laufer, 1997; Murakami & Alexopoulou, 2016), it remains unclear whether, say, metaphorical phrasal verbs that commonly take an 'infinitive' form such as 'put in' (rather than 'putting in') would be any more or less noticeable and acquirable than those that do not.

8.3.2.2 Specific NNS problems: English Grammatical Metaphoric Competence

Test 1-Phrasal Verbs-R and –P

The NNS response data for Test 1-Phrasal Verbs-R and –P revealed some group-wide issues. I will focus on one of these, the NNS tendency to select and produce 'on' or 'up' for receptive and productive items eliciting 'in'. However, it should be noted that similar issues can be found in responses to 'up', 'off', 'out' and 'down phrasal verbs too. Given the range of possible *incorrect*

Table 8.1 *Test 1-Phrasal Verbs-R: Particles Selected for Items Eliciting 'in' (Distractor Analysis)*

				NNSs				NSs				
No.	Test item	Option	Answer	HG	LG	Total (%)	Utility	HG	LG	Total (%)	Utility	Conceptual metaphor
6	I'll try to get _____ (do) an hour of reading before dinner.	**in**	✓	**3**	**1**	**8(14)**	**1**	**4**	**4**	**14(100)**	**1**	THE MIND IS A CONTAINER (Kurtyka, 2001)
		on	✗	8	14	36(64)	1	0	0	0(0)	0	
		with	✗	4	1	9(16)	-1	0	0	0(0)	0	
		out	✗	1	0	3(5)	-1	0	0	0(0)	0	
10	There's been an accident. We're still waiting for more news to come _____(arrive).	**in**	✓	**7**	**1**	**10(18)**	**1**	**4**	**4**	**14(100)**	**1**	RECEIVING INFORMATION IS AN OBJECT ENTERING (Kurtyka, 2001)
		up	✗	6	11	35(63)	1	0	0	0(0)	0	
		over	✗	2	3	9(16)	1	0	0	0(0)	0	
		on	✗	1	1	2(4)	0	0	0	0(0)	0	
15	I'm not asking you to put _____(contribute) too much time, just one or two hours a week.	**in**	✓	**15**	**6**	**36(64)**	**1**	**5**	**5**	**16(100)**	**1**	CONTRIBUTING TIME IS FILLING A CONTAINER (Kurtyka, 2001)
		on	✗	1	9	15(27)	1	0	0	0(0)	0	
		through	✗	0	0	3(5)	0	0	0	0(0)	0	
		up	✗	0	1	2(4)	1	0	0	0(0)	0	
18	With this new job I can bring _____(earn) enough money to pay my daughter's tuition fees	**in**	✓	**13**	**8**	**34(61)**	**1**	**5**	**5**	**16(100)**	**1**	POSSESSION IS CONTAINMENT (Neagu, 2007)
		up	✗	2	4	14(25)	1	0	0	0(0)	0	
		out	✗	0	2	5(9)	1	0	0	0(0)	0	
		over	✗	1	2	3(5)	1	0	0	0(0)	0	

Table 8.2 *Test 1-Phrasal Verbs-P: Particles Produced for Items Eliciting 'in'*

THE MIND IS A CONTAINER (Kurtyka, 2001)			
Item 6: 'get in' (productive)			
		Raw count (%)	
Particle	Answer	NNSs	NSs
on/on with	✗	23(41)	0(0)
in	✓	**10(18)**	**10(63)**
through	✓	**9(16)**	**3(19)**
down	✗	5(9)	1(6)
to	✗	2(4)	0(0)
over	✗	2(4)	0(0)
?	✗	2(4)	0(0)
off	✗	1(2)	0(0)
out	✗	1(2)	0(0)
up	✗	1(2)	0(0)
about	✗	0(0)	1(6)
Total		56(100)	15(100)

RECEIVED INFORMATION IS PHYSICAL ENTRY (Kurtyka, 2001)			
Item 10: 'come in' (productive)			
		Raw count (%)	
Particle	Answer	NNSs	NSs
up	✗	28(50)	1(6)
in	✓	**13(23)**	**8(50)**
out	✗	7(13)	2(13)
over	✗	4(7)	0(0)
on	✗	2(4)	0(0)
around	✗	1(2)	0(0)
through	✓	**1(2)**	**4(25)**
Total		56(100)	15(100)

EXPENDING ENERGY/CONTRIBUTING TIME IS FILLING A CONTAINER (Kurtyka, 2001)			
Item 15: 'put in' (productive)			
		Raw count (%)	
Particle	Answer	NNSs	NSs
in	✓	**26(46)**	**14(93)**
on	✗	16(29)	0(0)
up	✗	5(9)	0(0)
away	✗	2(4)	0(0)
across	✗	1(2)	0(0)
aside	✗	1(2)	0(0)
down	✗	1(2)	1(7)
forward	✗	1(2)	0(0)
into	✗	1(2)	0(0)
out	✗	1(2)	0(0)
over	✗	1(2)	0(0)
Total		56(100)	15(100)

POSSESSION IS CONTAINMENT (Neagu, 2007)			
Item 18: 'bring in' (productive)			
		Raw count (%)	
Particle	Answer	NNSs	NSs
in	✓	**27(48)**	**13(87)**
up	✗	13(23)	0(0)
on	✗	4(7)	0(0)
out	✗	3(5)	0(0)
about	✗	2(4)	0(0)
with	✗	2(4)	0(0)
along	✗	1(2)	0(0)
back	✗	1(2)	0(0)
down	✗	1(2)	0(0)
into	✗	1(2)	0(0)
?	✗	1(2)	0(0)
home	✓	**0(0)**	**2(13)**
Total		56(100)	15(100)

particles that could have been selected and produced for 'in' verbs, it is quite remarkable two emerged as systemic. Table 8.1 shows the receptive response data for NNSs and NSs (from the distractor analysis) for these items, the four options (*correct* answer marked '✓', distractors marked '✗'), each option's raw number (and percentage) of endorsements for NNSs and NSs (higher and lower groups, and total),[89] utility scores,[90] and (possible) conceptual metaphors engaged.

Table 8.2 shows produce responses and lists the different particles produced by NNSs and NSs, which of these were scored as '1' (*correct*) or '0' (*incorrect*) indicated by a '✓' or '✗',[91] and the raw number (and percentages) of NNS and NS productions[92] in the order most-to-least frequent for the NNSs.

For item 6 ('get <u>in</u>'), the number of *incorrect* NNS selections and productions of 'on' (64% and 41%) outweigh the number of *correct* selections and productions of 'in' (14% and 18%, with the 16% 'through' productions also scored as *correct*). For item 10 ('come <u>in</u>'), the number of *incorrect* NNS selections and productions of 'up' (63% and 50%) outweigh the number of *correct* selections and productions of 'in' (18% and 23%, with the 2% 'through' productions also scored as *correct*). For item 15 ('put in'), although a majority of NNSs selected and produced the *correct* answer 'in' (64% and 46%), the strongest distractor and the most common *incorrect* production was 'on' (27% and 29%). Similarly, for item 18 ('bring <u>in</u>'), although a majority of NNSs *correctly* selected and produced 'in' (61% and 48%), the most common distractor and *incorrect* production was 'up' (25% and 23%). By contrast, none of the distractors lured any of the NSs in the receptive mode, and although NSs produced a handful of *incorrect* particles, only one of these involved 'up' (item 10, 'come <u>in</u>'), one of the two problematic NNS particles for the phrasal verbs involving 'in'.

These findings show that the following interlanguage forms were salient in the NNS group data:

**I'll try to get <u>on</u> (do) an hour of reading before dinner.*

**There's been an accident. We're still waiting for more news to come <u>up</u> (arrive).*

**I'm not asking you to put <u>on</u> (contribute) too much time, just one or two hours a week.*

---

[89] Total counts are based on 56 NNSs and 14 NSs for group 1 since N3A had been deleted as an outlier (Appendix E), and 56 NNSs and 16 NSs for group 2, higher and lower group counts are based on the top and bottom 27% (approximately) of group 1 and group 2 NNSs and NSs.

[90] For *correct* answers: '1' = more HG than LG endorsements; '0' = equal HG and LG endorsements; '-1' = more LG than HG endorsements. For distractors: '1' = more LG than HG endorsements; '0' = equal HG and LG endorsements; '-1' = more HG than LG endorsements.

[91] The reader will note that for productive items 6, 10 and 18, it was necessary to score more than one particle as '1' (*correct*), or rather, for these items it would not have been fair to score equally acceptable particles as '0' (*incorrect*) simply because they were not the ones intended for elicitation.

[92] Raw numbers and percentages are based on 56 NNSs and 15 NSs from group 1, and 56 NNSs and 15 NSs from group 2 since N10B had been deleted as a participant outlier (Appendix E).

**With this new job I can bring up (earn) enough money to pay my daughter's tuition fees*

From a cognitive perspective, these data may be indicative of issues related to the acquisition of English conceptual metaphors, or suppression of Chinese ones. However, whether or not the issue really concerned the fact that NNSs had trouble conceptualising THE MIND, RECEIVING INFORMATION, CONTRIBUTING TIME, and POSSESSION in terms of CONTAINMENT (Kurtyka, 2001; Neagu, 2007) is a matter open to further research. Although interventions aimed at raising learners' awareness of possible conceptual metaphors suggested by prepositions may benefit some learners, such approaches comes with several warnings and disclaimers (MacArthur, 2010; Nacey, 2013). These implications are discussed in section 11.4.

Another arguably more viable interpretation of these interlanguage forms, is in terms of the NNSs inability to reject non-nativelike forms as being non-nativelike. This issue can be seen clearly in response data for the other 'excellent' marker of **English Grammatical Metaphoric Competence**, Test 3-Vehicle Acceptability-R.

Test 3-Vehicle Acceptability-R

For this test, the ten items designed to be highly acceptable were:

1) His **blood began to boil** as he started shouting (NNN p = 0.51, D = 0.35; NS p = 0.97, D = 0.11)
4) The whole theory **fell apart** (NNS p = 0.53, D = 0.87; NS p = 0.93, D = 0.11)
5) The project is **going ahead** as planned (NNS p =0.46, D = 0.71; NS p = 1.00, D = 0.00)
6) He couldn't **bottle his anger up** anymore so… (NNS p = 0.43, D = 0.35; NNS p = 0.90, D = 0.22)
7) It was an **attractive** proposal (NNS p = 0.49, D = 0.45; NS p = 0.93, D = 0.22)
8) The idea **holds up** in principle (NNS p = 0.50, D =0.48; NS p = 0.93, D = 0.22)
9) The drunken man was **repulsive** (NNS p = 0.38, D = 0.61; NNS p = 0.93, D = 0.22)
18) He told a **white** lie (NNS p = 0.69, D = 0.55; NS p = 1.00, D = 0)
19) She made a **firm** proposal to the client (NNS p = 0.66, D = 0.55; NS p = 0.92, D = 0.11)
20) He tried to **pull the wool over my eyes** (NNS p = 0.24, D = 0.45; NS p = 1.00, D = 0)

Item difficulty (p) and discriminability (D) indexes are shown in parenthesis. These can be compared with the eight lower acceptability items:

10) The theory was **the colour of brick** (NNS p = 0.22, D = 0.45; NS p = 0.93, D = 0.11)
12) Her hair had almost **arrived at** being grey (NNS p = 0.52, D = 0.42; NS p = 0.90, D = 0.33)
13) We **entered the front door** of the plan (NNS p = 0.23, D = 0.26; NS p = 0.87, D = 0.33)
15) Their similarities **jerked** them together (NNS p = 0.35, D = 0.45; NS p = 0.87, D = 0.33)
16) She **turned orange** as she started shouting at him (NNS p = 0.25, D = 0.39; NS p = 0.90, D = 0.22)
25) The comment **blunts** (NNS p = 0.04, D = 0.13; NS p = 0.90, D = 0.22)
26) We asked for **a called day** at 6pm. (NNS p = 0.19, D = 0.32; NS p = 0.93, D = 0.11)
28) We asked for **a show of the ropes** (NNS p = 0.25, D = 0.26; NS p = 0.87, D = 0.22)

A comparison of the item difficulty scores shows that for the 10 higher acceptability items, the NNS values had *M* = .49 (*SD* = .13) and NS values *M* = .95 (*SD* = .04), whereas for eight lower acceptability items, the NNS values had *M* = .26 (*SD* = .14) and the NS values *M* = .90 (*SD* = .03). Comparing these differences, the effect size is greater for lower than higher acceptability items (Cohen's *d* = -6.494 vs -4.938). These findings tentatively suggest that there is a greater gap between NNSs and NSs for the skill of rejecting non-nativelike Vehicle terms as unacceptable, rather than for accepting nativelike ones as acceptable.

The greater differences between NNSs and NSs for lower acceptability items speaks to the logical inference that the former have had inadequate exposure to negative evidence, namely examples of what native speakers tend not to say (Low, 1988). Because none of the NNSs in the present study had acquired English in an immersion setting before the critical period (section 2.4.2), for them, the task of developing sensitivity to nativelike and non-nativelike word combinations akin to that of NSs is likely to be insurmountable (Foster et al., 2014). However, by living and studying in an immersion setting (the UK), these participants could hope to make substantial gains in their receptive sensitivity (albeit not to nativelike level). Concerning productive knowledge, past research would predict that the longer that the NNSs spent living in the UK, the more diverse their lexis would become, and the more their word combinations would resemble those of NSs (Foster & Tavakoli, 2009); much of this is likely to involve metaphor. While comprehensive investigation into these issues requires a separate study in itself, it was possible to explore the extent of any relationship between age of starting to learn English, length of stay in the UK and receptive and productive MC in the present study's NNSs. This analysis forms part of the next chapter.

8.3.3 L2 English Illocutionary Metaphor Production, English Metaphor Language Play, and English Topic/Vehicle Acceptability: The same yet different…but not deficient

The NNSs and NSs were statistically equivalent in terms of their **English Illocutionary Metaphor Production**, **English Metaphor Language Play**, and **English Topic/Vehicle Acceptability**. But were they really indistinguishable on these factors?

The notion that NNSs can be 'different' from NSs, but not 'deficient' is an issue that has been raised in the present study with regard to ELF (section 2.4.4), but also one that lies at the heart of other research agendas, for instance involving heritage languages (e.g., Bayram et al., 2017; Kupisch & Rothman, 2016). The discussion below aims to highlight some areas of difference-but-not-deficiency in the NNSs compared with the NSs. Due to space limitations, it is a very general overview, giving only a flavour of the qualitative differences between NNSs and NSs, and possible areas that might form a basis for further research.

Concerning **English Metaphor Language Play**, several qualitative NNS and NS differences were observed in responses for Test 8-Idiom Extension-P. For item 5 'after her email the ball is in my court…the problem is ____', most '2' scoring NNSs (13 out of 15, 87%) reproduced the lexical item 'ball' in their response, compared with only a fifth of NSs (3 out of 15, 20%). The prevalence of 'ball' in the NNS response data, as well as the comments of two NNS think aloud participants point to potential strategy use concerning mental (or verbalised) repetition and/or 'circling' of this word to help facilitate extension of the idiom (Littlemore & Low, 2006a). When a specific sport was mentioned for this item, 40% of NSs (6 out of 15) responses refer to a racquet sport (or one involving a net), whereas the NNSs preferred to draw on throwing sports, football, and baseball. For item 4 'don't worry…go out and break a leg. In fact, go out and_____', most '2' scoring NNSs and NSs responded by pluralising 'leg' (e.g., 'break both legs', 'break two legs'), however, the NSs seemed to prefer the alliterative collocation 'break both' (10 out of 15, 66%), whereas the NNSs favoured 'break two' (4 out of 9, 44%). Although one may postulate numerous explanations for these differences (e.g., more stable NS lexical networks, the British obsession with tennis), a further study involving, for example, introspective methods would be needed to properly tap into the reasons why NNSs and NSs extended idioms in the ways they did.

For **English Illocutionary Metaphor Production**, NNS and NS differences were also found. For Test 6-Heuristic-P item 11 'the heart functions like ____', most '2' scoring NSs and NNSs mentioned the lexical item 'pump'. While the NSs also responded with common, household objects/entities like 'clock' (N10A) and 'engine of a car' (N13A, N14A), the NNSs produced more imaginative, unusual similes such as "perpetual motion machine" (40A), "the power station of the body" (43A), and "drum beats that keep a band alive" (48A).

For **English Topic/Vehicle Acceptability**, Test 4-Topic/Vehicle-R item 4, 'the main argument is the _____ of the essay' (best answer = 'meat'), the distractor analysis revealed that the difference between the NNSs and NSs seems to be primarily to do with how they perceived the acceptability of 'meat' and 'bread' as analogies of 'a main argument'. While 'meat' was rated as the most acceptable Vehicle by both NNSs and NSs, groups differed widely on the extent to which they found this answer acceptable, the NSs giving it a mean group rating of 97.33% acceptable, the NNSs finding it 59.47% acceptable. The distractors revealed numerous NNS and NS differences. All distractors, 'bread', 'rice' and 'pasta' received distinctly higher NNS than NS ratings, suggesting again that NSSs struggled to reject non-nativelike forms as such. These differences are also likely to be indicative of the culturally loaded metaphorical meanings that such foodstuffs have in English and Chinese (Littlemore & Low, 2006a).

8.3.4 L2 phraseological proficiency vs conceptual fluency

The finding that NNSs and NSs differed with regard to a grammatical aspect of metaphoric competence but not others largely aligns with Johnson and Rosano's (1993) finding that NNSs of English (mostly L1 Mandarin) and NSs differed on de-contextualised measures of vocabulary and verbal analogies, but not in terms of the complexity and fluency of their metaphor interpretations when individual, constituent words were known. The fact that the present study's NNSs produced such a rich dataset is evidence that as a group, they were remarkably capable of engaging with the concepts and layers of meaning presented to them (Danesi, 1992, 1995). Importantly, a decision was taken at the start not to score productive responses for grammatical accuracy (section 4.3.3). This dimension was therefore removed from all analyses. However, had it been included, the pervasiveness of grammatical inaccuracy in the NNS productions (evident in some of the examples cited above, and those below) would likely have revealed that even at higher levels, phraseological proficiency lags behind conceptual fluency (Philip, 2010).

For instance, when responding to Test 5-Topic Transition-P item 2, NNSs realised the phrase 'when in Rome do as the Roman's do' in several (grammatically inaccurate) ways: 'when in Rome, do as Romes do.' (10B), 'do in Rome as Rome does.' (16B), and 'do in Romes as Rome does' (32B). Since test taker responses were not scored for grammatical accuracy, these productions were all scored as '2' (*correct*), since they were all recognisable attempts at the common proverb. The grammatical inaccuracies seem to suggest that these learners had processed the individual constituents of this formulaic sequence separately, consequently misapplying grammatical rules such as plural, third person and/or possessive –s (depending on which meanings were intended) This finding is unsurprising given the typological distance between English and Chinese and the substantial body of literature the predicts problems with these forms for L1 Chinese learners of English (Murakami & Alexopoulou, 2016). This example serves to show that while NNSs and NSs often both demonstrated knowledge of a particular metaphor, phrase, proverb, saying and so on, this quality of this knowledge was not always the same.

8.4 Chapter summary

In this chapter, the key findings of **Analysis 2: Metaphoric and other (sub)competences uncovered** were discussed in relation to previous metaphoric competence and other research. Concerning the third research question, the present study showed that several factors, representing various (sub)competences do underlie the NNS data (i.e., L2 MC, vocabulary knowledge and general proficiency test scores).

Although this study is believed to be the first EFA approach to L2 MC, it sought to

improve on basic methodological issues identified in similar past L1 MC research (e.g., inappropriate factor extraction technique, reporting measures of model adequacy, consistent use of and empirical basis for qualitative descriptors 'high', 'strong', 'stable', etc.). In combination with this research (Beaty & Silvia, 2013; Mashal & Kasirer, 2012; H. R. Pollio, 1977; H. R. Pollio & Smith, 1980; Silvia & Beaty, 2012), the present study's findings suggested that regardless of whether one is exploring MC in the L1 or the L2, a four-to-six factor solution is likely to be the most parsimonious and interpretable from between 20 and 30 metaphoric competence variables. One major issue concerned the disparate number of factors suggested for various retention criteria (anywhere from one to 12). While it was argued that a six-factor solution was the most parsimonious and interpretable for the NNS EFA, these considerations showed that further investigation on approaches to factor retention in L2 MC research is needed.

For the NNS EFA, the six-factor solution interpreted led to two main conclusions. First, L2 MC, **English Vocabulary Size** and **English General Comprehension** appear to be distinct constructs, evidenced by the emergence of these two (largely) non-MC factors. Second, L2 MC itself appears to comprise the (sub)competences **English Grammatical Metaphoric Competence**, **English Illocutionary Metaphor Production**, **English Metaphor Language Play**, and **English Topic/Vehicle Acceptability**. Although there is no past L2 MC EFA to compare these findings with, the fact that **English Grammatical Metaphoric Competence** appears to mostly concern 'conventional' metaphor, whereas **English Metaphor Language Play** concerns 'creative' metaphor finds a(n albeit imperfect) parallel in L1 MC research (Beaty & Silvia, 2013; Mashal & Kasirer, 2012; H. R. Pollio, 1977; H. R. Pollio & Smith, 1980; Silvia & Beaty, 2012).

Although the homogeneity of **English Grammatical Metaphoric Competence** and **English Illocutionary Metaphor Production** with regard to Low (1988) and Littlemore and Low's (2006a) framework components vouched for the conceptual independence of these (sub)constructs, the mixture of variables loading on **English Topic/Vehicle Acceptability** and **English Metaphor Language Play**, and the fact that some factors were predominantly receptive or productive suggests that the authors' frameworks might be further scrutinised, and if necessary, revised. To this end, the present study has produced several informal hypotheses that require testing in further research.

Concerning the fourth research question, the EFA of the NNS+NS data showed that the four MC factors (**English Grammatical Metaphoric Competence**, **English Illocutionary Metaphor Production**, **English Metaphor Language Play**, and **English Topic/Vehicle Acceptability**) could be found in both the NNS and NNS+NS data sets (with some differences between loading variables), and that NNSs (L1 Chinese) and NSs (L1 English) differed in their overall MC (i.e., for all factors) but only in terms of their **English Grammatical Metaphoric**

Competence, when individual factors are considered. Some possible reasons why **English Grammatical Metaphoric Competence** may be the hardest aspect of L2 metaphoric competence to acquire were discussed.

Phrasal verbs, a major aspect of this (sub)competence, seem to have posed various problems. However, the fact that NNSs of English from L1s such as Dutch (which has an equivalent to phrasal verbs) also experience problems with these forms, suggests that the difficulties experienced by the present study's NNSs (L1 Chinese) may be somewhat universal, even though the specific reasons underlying phrasal verb difficulty may vary from one L1 group to the next (Kellerman, 1983; Liao & Fukuya, 2004). A probe into the potential relationship between frequency of phrasal verb forms and receptive and productive item difficulty showed no correlations. While this analysis is limited for several reasons (e.g., frequencies concerned any corpus hits for phrasal verbs forms, rather than the specific form-meaning mapping used in test items), it showed that **English Grammatical Metaphoric Competence** (as measured) is likely to be more strongly affected by *non-frequency* factors (e.g., perceived saliency of forms, representation in the L1, degree of metaphoricity).

A comprehensive investigation into NNS and NS differences in **English Grammatical Metaphoric Competence** was beyond the scope of the present study, however, a close analysis of one issue showed that NNSs had a tendency to select and produce 'on' or 'up' for phrasal verbs eliciting 'in'. While a conceptual metaphor based account of these differences is speculative at best, one can reach a more productive position by comparing these data with the NNSs failure to reject unacceptable Vehicle terms as such for Test 3-Vehicle Acceptability-R (another 'excellent' marker of **English Grammatical Metaphoric Competence**). Taken together, these issues seem to point to the conclusion that the NNSs require yet more evidence of what native speakers tend not to say (Low, 1988) if they are to improve sensitivity to nativelike and non-nativelike uses of metaphor. Given the findings of past research (e.g., Foster et al., 2014; Foster & Tavakoli, 2009), one could predict that spending time in the UK (an L2 rich environment) would bring the greatest gains for productive rather than receptive **English Grammatical Metaphoric Competence** for these speakers (although not to nativelike levels). However, this prediction would need to take into account a range of other potentially influential variables (e.g., phonological short-term memory).

The NNSs and NSs were statistically indistinguishable in terms of their **English Illocutionary Metaphor Production**, **English Metaphor Language Play**, and **English Topic/Vehicle Acceptability**. This was intriguing and warranted exploration into whether there were in fact systemic qualitative differences. This discussion highlighted differences in lexical items, collocations and concepts invoked, as well as apparent strategy use. Importantly, it was argued that these NNS and NS differences should not be seen as NNS deficiencies. Rather, these

points might serve as a basis for further research into how NNSs and NSs behave with regard to these (statistically equivalent) areas of MC.

Finally, although the scoring criteria for productive responses concerned 'meaning quality' only, (i.e., not 'grammatical accuracy'), some NNS and NS differences in phraseological proficiency were discussed (Philip, 2010).

Thus far, the reliability of the MC Test Battery, basic descriptive statistics, and Metaphoric and other (sub)competences underlying the data have been investigated. In the next two chapters, I conclude the analyses by exploring the relationships between L2 metaphoric competence, vocabulary knowledge, general proficiency and factors related to age and time.

Chapter 9: Analysis 3 - Relationships between L2 metaphoric competence, vocabulary knowledge, general proficiency, age of starting to learn English and time spent living in the UK

9.1 Introduction

In the first part of this chapter, three sets of regression analyses that explore the ability of L2 vocabulary knowledge, L2 proficiency, age of starting to learn English, and time spent living in the UK to predict L2 metaphoric competence are reported. These analyses help answer research question 5. The second part of the chapter reports changes in the correlations between L2 receptive and productive metaphoric competence at different levels of L2 proficiency. These analyses help answer research questions 6.

9.2 Regression 1: L2 metaphoric competence predicted by L2 vocabulary knowledge

9.2.1 Data screening

Multiple regression involves several assumptions. Consequently, the NNS data were first screened using Pallant's checklist (2013, pp. 156-157). Screening pertains to this checklist unless otherwise stated. Presented below is a summary of this process. In this chapter, the terms 'IV' and 'predictor' are synonymous, as are 'DV' and 'criterion variable'.

The process of data screening revealed that:

- Two scores (35A's MC-R score and 25B's VYesNo score) were outliers,[93] however, since these were plausible values and differences between the means and trimmed means were negligible,[94] they were allowed to remain;
- The sample size (109, 112, 109, 111 and 111 for MC-R, MC-P, MC-R&P, VYesNo, and WAT respectively) was sufficient by Tabachnick and Fidell's (2013), Field's (2013),

[93] Since multiple regression is particularly sensitive to outliers, a stricter method for outlier detection (deletion of cases lying more than 1.5. box-lengths from the edge of the box) was implemented than in the EFA.

[94] The mean-to-trimmed mean ratio was 1.0007:1 for metaphoric competence receptive and 0.995:1 for the VYesNo. These ratios are very similar to 1.0033:1, Pallant's example for a situation in which outliers were not deleted (2013, p. 67).

Cohen's (1988) and Khamis and Kepler's (2010) criteria given that in all cases, at least one predictor had R^2 and beta (β) values above 0.3;

- Data met the assumptions for multicollinearity, evidenced by the absence of high correlations (Pallant suggests $r > .7$ could be problematic), and the fact that the Tolerance statistic and VIF values for the two IVs (VYesNo and WAT) were 0.695 and 1.439 respectively, thus well above and below the suggested .1 and 10;
- Singularity was not an issue, since no analysis involved IVs comprised from DVs or vice versa;[95]
- There were no violations of normality, linearity, homoscedasticity and independence of residuals in the data evidenced by the Normal Probability Plots (P-P) of the Regression Standardised Residual and Scatterplots;
- Post-hoc checks revealed no outliers, evidenced by the fact that the standardised residuals (as displayed in the scatterplot) showed no cases greater than +3.3 or less than -3.3, and because the largest Mahalanobis distance was 8.73 (participant 50B), which is below 13.82, the chi-square critical value when the degrees of freedom = 2 [independent variables] ($p < .001$).

9.2.2 Results

9.2.2.1 Model 1: MC-R predicted by VYesNo and WAT

How well did VYesNo and WAT scores predict MC-R scores, and which was the best predictor?

The total variance in MC-R scores explained by Model 1 as a whole was 35% ($R^2 = 0.350$), $F(2,105) = 28.256$, $p < .001$. Table 9.1 shows that the WAT had a slightly larger beta coefficient than VYesNo, 0.348 compared with 0.323, indicating that it made a slightly stronger unique contribution to explaining variance in MC-R scores, when all other variance in the model was controlled for. Both beta coefficients were found to be statistically significant ($p < .001$ and $p = .001$ respectively):

Table 9.1 *Model 1: Coefficients*[a]

Model		Unstandardized Coefficients		Stand. Coeffs.			95% Confidence interval for *B*		Correlations		
		B	*SE*	Beta (β)	*t*	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-17.530	12.128		-1.445	.151	-41.578	6.519			
	VYesNo	.003	.001	.323	3.424	.001	.001	.005	.516	.317	.269
	WAT	.419	.114	.348	3.686	.000	.194	.644	.527	.339	.290

[a] Dependent variable: MC-R.

[95] Although MC-R&P scores were composites of MC-R and MC-P scores, these variables were not used in the same analysis.

Thus, for every one unit increase in WAT scores (i.e., every associate recognised) or VYesNo scores (i.e., every new word recognised), an increase in MC-R of .346 and .323 *SD* units respectively could be expected, controlling for the effect of the other predictor.

By squaring the part correlations, it was revealed that 8.41% of variance in MC-R scores was uniquely explained by the WAT, with any overlap or shared variance partialled out or removed, and 7.24% by VYesNo.

In summary, the model, which included controls of VYesNo and WAT, explained 35% of the variance in MC-R scores. Of these two variables, the WAT made a slightly larger contribution (β = 0.348) than VYesNo (β = 0.323); both contributions were statistically significant at the .01 level.

9.2.2.2 Model 2: MC-P predicted by VYesNo and WAT

How well did VYesNo and WAT scores predict MC-P scores, and which was the best predictor?

The total variance in MC-P scores explained by Model 2 as a whole was 39.4% (R^2 = 0.394), $F(2,107) = 34.774$, $p < .001$. Table 9.2 shows that the WAT had a larger beta coefficient, 0.526, suggesting that this variable made the strongest unique contribution to explaining the variance in the MC-P scores, when all other variance in the model was controlled for. The beta value for VYesNo was lower at 0.159. While the beta coefficient for the WAT was statistically significant at the .01 level, the coefficient for VYesNo was not significant at the .05 level (but was significant at the .1 level):

Table 9.2 *Model 2: Coefficients*[a]

Model		Unstandardized Coefficients		Stand. Coeffs.			95% Confidence interval for *B*		Correlations		
		B	*SE*	*Beta* (β)	*t*	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-75.284	15.751		-4.780	.000	-106.509	-44.060			
	VYesNo	.002	.001	.159	1.761	.081	.000	.005	.449	.168	.133
	WAT	.859	.148	.526	5.822	.000	.567	1.152	.613	.490	.438

[a] Dependent variable: MC-P.

In other words, for every one unit increase in WAT scores (i.e., every associate recognised) or VYesNo scores (i.e., every new word recognised), an increase in MC-P of .526 and .159 *SD* units respectively could be expected, controlling for the effect of the other predictor.

By squaring the part correlations, it was revealed that 19.18% of variance in MC-P was uniquely explained by the WAT, with any overlap or shared variance partialled out or removed, and 1.77% by VYesNo.

In summary, the model, which included controls of VYesNo and WAT, explained 39.4%

of the variance in MC-P scores. Of these two variables, WAT made the largest contribution (β = 0.526), significant at the .01 level, whilst VYesNo made a much smaller contribution (β = 0.159), not significant at the .05 level (but significant at the .1 level).

9.2.2.3 Model 3: MC-R&P predicted by VYesNo and WAT

How well do VYesNo and WAT scores predict MC-R&P scores, and which was the best predictor?

The total variance in MC-R&P scores explained by Model 3 as a whole was 43.6% (R^2 = 0.436), $F(2,105) = 40.564$, $p < .001$. Table 9.3 shows that the WAT had a larger beta coefficient, 0.494, suggesting that this variable made the strongest unique contribution to explaining the variance in MC-R&P scores, when all other variance in the model was controlled for. The beta value for VYesNo was slightly lower at 0.243. Both beta coefficients were found to be statistically significant ($p < .001$ and $p = .007$ respectively).

Table 9.3 *Model 3: Coefficients*[a]

Model		Unstandardized Coefficients		Standard. Coeffs.			95% Confidence interval for *B*		Correlations		
		B	*SE*	*Beta* (β)	*t*	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-43.057	11.828		-3.640	.000	-66.509	-19.604			
	VYesNo	.003	.001	.243	2.760	.007	.001	.005	.516	.260	.202
	WAT	.623	.111	.494	5.621	.000	.403	.843	.628	.481	.412

[a] Dependent variable: MC-R&P.

In other words, for every one unit increase in WAT scores (i.e., every associate recognised) or VYesNo scores (i.e., every new word recognised), an increase in MC-R&P of .494 and .243 *SD* units respectively could be expected, controlling for the effect of the other predictor.

By squaring the part correlations, it was revealed that 16.97% of variance in MC-R&P was uniquely explained by the WAT, with any overlap or shared variance partialled out or removed, and 4.08% by VYesNo.

In summary, the model, which included controls of the WAT and VYesNo, explained 43.6% of the variance in MC-R&P scores. Of these two variables, WAT made the largest contribution (β = 0.494) whilst VYesNo also made contribution (β = 0.243); both were significant at the .01 level.

9.2.2.4 Magnitude of predictive power: Hierarchical regression

The predictive values of the independent variables in models 1-3 were explored further using hierarchical regressions aimed at investigating the magnitude of R^2 changes. With the WAT (best predictor) entered into the model at the first step and VYesNo (second best predictor) entered at the second step, the R^2 change was .073 (F change = 11.727, $p < .01$) for MC-R; .018 (F change

= 3.100, p = .81) for MC-P; and .041 (F change = 7.620, p < .01) for MC-R&P. With VYesNo (second best predictor) entered into the model at the first step and WAT (best predictor) entered at the second step, the R^2 change was .084 (F change = 13.588, p < .01) for MC-R; .192 (F change = 33.893, p < .01) for MC-P; and .170 (F change = 31.600, p < .01) for MC-R&P.

In other words, VYesNo (second best predictor) provided an additional 7.3%, 1.8% and 4.1% of the criterion (DV) variance over and above the WAT (best predictor) for MC-R, MC-P, MC-R&P respectively, whereas the WAT (best predictor) provided an additional 8.4%, 19.2% and 17% of the criterion (DV) variance of these variables over and above VYesNo (second best predictor). All changes were statistically significant, suggesting that a combination of the two variables offered more explanatory power than any one in isolation.

9.2.2.5 Summary

In summary, in a combined model, VYesNo and WAT were able to significantly predict 35%, 39.4% and 43.6% of the total variance in MC-R, MC-P and MC-R&P scores respectively. Both VYesNo and WAT had significant, predictive power, with the WAT proving superior for all modes. These data are summarised in Table 9.4:

Table 9.4 *Regression 1: R^2 Values for Individual and Combined Predictors of MC*

Model	Criterion variable (DV)	Predictor variable (IV)	R^2
1	MC-R	VYesNo	0.266**
		WAT	0.278**
		Combined model	0.350**
2	MC-P	VYesNo	0.202**
		WAT	0.376**
		Combined model	0.394**
3	MC-R&P	VYesNo	0.266**
		WAT	0.394**
		Combined model	0.436**

**significant at the .01 level.

9.3 Regression 2: L2 metaphoric competence predicted by L2 general proficiency components

9.3.1 Data screening

A total of 27 outliers, exceeding more than 1.5 box lengths from the edge of boxes were detected in the L2 proficiency component variables. These included 44A and 29B for OOPT Use of English; 56A for IELTS Reading; 2A, 35A, 46A, 50A, 8B, 9B, 22B, 23B, 38B, 47B (low outliers) and 28A, 43A, 56A, 12B, 43B, 44B, 46B, 56B (high score outliers) for IELTS Writing; and 4A, 35A, 45A, 12B, 14B, 17B for IELTS Speaking. However, since all these scores fell within plausible ranges, and the means and trimmed means were very similar in all cases, no scores were deleted. Further data

screening revealed no issues with sample size, multicollinearity, singularity, normality, linearity, homoscedasticity, independence of residuals and post-hoc outlier checks. Due to limited space, these are not presented but are available upon request.

9.3.2 Results

9.3.2.1 Model 4: MC-R predicted by OOPT and IELTS components

How well did OOPT Use of English, OOPT Listening, IELTS Reading, IELTS Writing, IELTS Speaking and IELTS Listening scores predict MC-R scores, and which was the best predictor?

The total variance in MC-R scores explained by Model 4 as a whole was 45.2% (R^2 = 0.452), $F(6,101) = 13.899$, $p < .001$. To answer this question, the six independent variables included in the model were evaluated. The results (Table 9.5) showed that the OOPT Listening had a largest beta coefficient, 0.299 (p = .001), suggesting that this variable made the strongest unique contribution to explaining the variance in the MC-R scores when all other variance in the model is controlled for. OOPT Use of English also made a significant contribution (β = 0.243, p = .008), whereas the IELTS strands made lower, non-significant contributions.

Table 9.5 *Model 4: Coefficients*[a]

Model		Unstandardized Coefficients		Stand. Coeffs.			95% Confidence interval for *B*		Correlations		
		B	*SE*	*Beta* (β)	*t*	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-25.243	13.770		-1.833	.070	-52.558	2.072			
	OOPT Use of Eng.	.204	.075	.243	2.725	.008	.056	.353	.521	.262	.201
	OOPT Listening	.217	.066	.299	3.296	.001	.086	.348	.537	.312	.243
	IELTS Reading	1.785	1.319	.122	1.354	.179	-.831	4.400	.436	.133	.100
	IELTS Writing	2.843	2.197	.107	1.294	.199	-1.516	7.201	.299	.128	.095
	IELTS Speaking	2.033	2.014	.088	1.009	.315	-1.962	6.029	.347	.100	.074
	IELTS Listening	1.455	1.262	.108	1.153	.252	-1.048	3.959	.452	.114	.085

[a] Dependent variable: MC-R.

In other words, for every one unit increase in OOPT Listening or OOPT Use of English scores (i.e., every mark gained), an increase in MC-R of .299 and .243 *SD* units respectively could be expected, controlling for the effect of the other predictors.

By squaring the part correlations, it was revealed that 5.9% of variance in MC-R was uniquely explained by OOPT Listening and 4% by OOPT Use of English, with any overlap or shared variance partialled out or removed, and lower amounts by the other predictors (IVs).

In summary, the model, which includes controls of the OOPT and IELTS components, explained 45.2% of the variance in MC-R scores. Of these six variables, the OOPT Listening made the largest contribution (β = 0.299) whilst OOPT Use of English also made a contribution (β = 0.243); both were statistically significant at the .01 level.

9.3.2.2 Model 5: MC-P predicted by OOPT and IELTS components

How well did OOPT Use of English, OOPT Listening, IELTS Reading, IELTS Writing, IELTS Speaking and IELTS Listening scores predict MC-P scores, and which was the best predictor?

The total variance in MC-P scores explained by Model 5 as a whole was 37.1% (R^2 = 0.371), $F(6,104) = 10.229$, $p < .001$. Table 9.6 shows that the OOPT Use of English had the largest beta coefficient, 0.310 (p = .001), suggesting that this variable made the strongest unique contribution to explaining the variance in the MC-P scores when all other variance in the model was controlled for. IELTS Listening and OOPT Listening also made modest contributions, although these were not significant at the .05 level (but were significant at the .1 level).

Table 9.6 *Model 5: Coefficients*[a]

Model		Unstandardized Coefficients		Stand. Coeffs.			95% Confidence interval for *B*		Correlations		
		B	*SE*	*Beta* (β)	*t*	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-38.871	19.742		-1.969	.052	-78.021	.279			
	OOPT Use of Eng.	.354	.107	.310	3.296	.001	.141	.567	.512	.308	.256
	OOPT Listening	.181	.094	.184	1.914	.058	-.007	.368	.452	.184	.149
	IELTS Reading	.820	1.891	.041	.434	.665	-2.929	4.569	.347	.043	.034
	IELTS Writing	.399	3.150	.011	.127	.899	-5.848	6.646	.198	.012	.010
	IELTS Speaking	2.597	2.888	.083	.899	.371	-3.130	8.324	.314	.088	.070
	IELTS Listening	3.576	1.809	.196	1.976	.051	-.012	7.164	.443	.190	.154

[a] Dependent variable: MC-P.

In other words, for every one unit increase in OOPT Use of English, IELTS Listening and OOPT Listening scores (i.e., every mark gained), an increase in MC-P of .310, .196 and .184 *SD* units respectively could be expected, controlling for the effect of the other predictors.

By squaring the part correlations, it was revealed that 6.6% of variance in MC-P was uniquely explained by OOPT Use of English, with any overlap or shared variance partialled out or removed, and around half that by IELTS Listening and OOPT Listening.

In summary, the model, which included controls of the OOPT and IELTS components, explained 37.1% of the variance in MC-P scores. Of these six variables, the OOPT Use of English made the largest contribution (β = 0.310), significant at the .01 level, whilst IELTS Listening and OOPT Listening also made modest contributions, although these were not significant at the .05 level (but were significant at the .1 level).

9.3.2.3 Model 6: MC-R&P predicted by OOPT and IELTS components

How well did OOPT Use of English, OOPT Listening, IELTS Reading, IELTS Writing, IELTS Speaking and IELTS Listening scores predict MC-R&P scores, and which was the best predictor?

The total variance in MC-R&P scores explained by Model 6 as a whole was 46.5% (R^2 = 0.465), $F(6,101) = 14.652$, $p < .001$. Table 9.7 shows that the OOPT Use of English had a larger beta

coefficient, 0.309 (p = .001), suggesting that this variable made the strongest unique contribution to explaining the variance in the MC-R&P scores when all other variance in the model was controlled for. OOPT Listening also made a substantial contribution (β = 0.259, p = .005), significant at the .01 level. IELTS Listening made a smaller contribution, although this was not significant at the .05 level (but was significant at the .1 level).

Table 9.7 *Model 6: Coefficients*[a]

Model		Unstand. Coeffs.		Stand. Coeffs.			95% CI for B		Correlations		
		B	*SE*	*Beta* (β)	*t*	Sig.	Lower	Upper	Zero-order	Partial	Part
1	(Constant)	-27.173	14.242		-1.908	.059	-55.425	1.079			
	OOPT Use of Eng.	.272	.077	.309	3.515	.001	.119	.426	.558	.330	.256
	OOPT Listening	.196	.068	.259	2.883	.005	.061	.331	.533	.276	.210
	IELTS Reading	1.078	1.364	.071	.790	.431	-1.628	3.783	.408	.078	.057
	IELTS Writing	2.238	2.273	.080	.985	.327	-2.270	6.746	.270	.098	.072
	IELTS Speaking	1.338	2.083	.055	.642	.522	-2.795	5.471	.329	.064	.047
	IELTS Listening	2.515	1.305	.179	1.927	.057	-.074	5.105	.481	.188	.140

[a] Dependent variable: MC-R&P.

In other words, for every one unit increase in OOPT Use of English, OOPT Listening and IELTS Listening scores (i.e., every mark gained), an increase in MC-R&P of .309, .259 and .179 *SD* units respectively could be expected, controlling for the effect of the other predictors.

By squaring the part correlations, it was revealed that 6.55% of variance in MC-R&P was uniquely explained by OOPT Use of English, with any overlap or shared variance partialled out or removed, and much lower amounts by the other predictors (IVs).

In summary, the model, which included controls of the OOPT and IELTS components, explained 46.5% of the variance in MC-R&P scores. Of these six variables, the OOPT Use of English made the largest contribution (β = 0.309) whilst OOPT Listening also made a contribution (β = 0.259); both were significant at the .01 level. IELTS Listening made a smaller contribution that was not significant at the .05 level (but was significant at the .1 level).

9.3.2.4 Magnitude of predictive power: Hierarchical regression

The predictive values of the independent variables were explored further using hierarchical regressions aimed at investigating the magnitude of R^2 changes. For MC-R, with OOPT Listening (best predictor) entered into the model at the first step and the other variables entered at the second step, the R^2 change was .164 (F change = 6.038, p < .01) for MC-R. With OOPT Use of English (second best predictor) entered into the model at the first step and the other variables entered at the second step, the R^2 change was .180 (F change = 6.652, p < .01). These data indicate that the rest of the variables provided an additional 16.4% of the criterion (DV) variance over and above the OOPT Listening (best predictor), and an additional 18% over and above the OOPT Use of English (second best predictor).

For MC-P, with the OOPT Use of English (best predictor) entered into the model at the first step and the other variables at the second, the R^2 change was .109 (*F* change = 3.607, *p* < .01) for MC-P. With IELTS Listening (second best predictor) entered into the model at the first step and the other variables entered at the second step, the R^2 change was .175 (*F* change = 5.772, *p* < .01). These data indicate that the rest of the variables provided an additional 10.9% of the criterion (DV) variance over and above the OOPT Use of English (best predictor), and an additional 17.5% over and above the IELTS Listening (second best predictor).

For MC-R&P, with the OOPT Use of English (best predictor) entered into the model at the first step and the other variables at the second, the R^2 change was .154 (*F* change = 5.814, *p* < .01). With OOPT Listening (second best predictor) entered into the model at the first step and the other variables entered at the second step, the R^2 change was .181 (*F* change = 6.847, *p* < .01). These data indicate that the rest of the variables provided an additional 15.4% of the criterion (DV) variance over and above the OOPT Use of English (best predictor), and an additional 18.1% over and above the OOPT Listening (second best predictor).

9.3.2.5 Summary

In summary, the OOPT and IELTs components were able to significantly predict 45.2%, 37.1% and 46.5% of the total variance in MC-R, MC-P and MC-R&P scores respectively. As individual predictors, the OOPT Listening was the best predictor of MC-R, whereas the OOPT Use of English was the best predictor of MC-P and MC-R&P.

9.4 Regression 3: L2 metaphoric competence predicted by L2 vocabulary knowledge, L2 general proficiency (overall), age of starting to learn English and time spent living in the UK

9.4.1 Data screening

Reported age of starting to learn English ranged from 3 to 18 years. While participant 36A, who reported starting to learn English at 18 years old, extended more than 1.5 box lengths from the edge of the box (i.e., was not a highly extreme outlier), the fact that in 2003, when 36A was 10 years old, English was introduced in China as compulsory subject in from Primary Three (i.e., age 8) (Qi, 2016) makes it unlikely that he would have been able to avoid English tuition until the age of 18. Since the next highest age of starting to learn English was 13 years, reported by eleven participants, 36A's data were removed from this variable.

Time spent living in the UK was reported in months, and ranged from less than 1 month

to 70 months. Two participants already had missing data for this variable (46A, 52A).[96] The box plots revealed two participants, 42B and 33B, who had lived in the UK for 70 and 48 months respectively, and who extended more than 3 box lengths from the edge of the box. These values were extreme cases, and resulted in quite a substantial difference between the mean and trimmed mean (1:0.78) and so were removed from this variable. Exceeding more than 1.5 box-lengths from the edge of the box were: 3A, 14A and 18A (36 months); 2B (32 months); 9B (27 months); and 17A, 20A, 27A and 43A (24 months). An initial regression, with these participants remaining was run, but revealed that one participant who had lived in the UK for 36 months (14A) had a Mahalanobis distance exceeding 22.458, the chi-square for when the degree of freedom = 6 [IVs] ($p < .001$). Because of this, all participants who had lived in the UK for 36 months (3A, 14A, 18A) were removed, and the analysis rerun. The highest Mahalanobis distance, 17.401 for participant 12B, was now below the critical cut-off. With no remaining cases exceeding 3 box lengths, this participant (and all others) was allowed to remain.

Further data screening again revealed no issues with sample size, multicollinearity, singularity, normality, linearity, homoscedasticity, independence of residuals and post-hoc outlier checks. Due to limited space, these are not presented but are available upon request.

9.4.2 Results

9.4.2.1 Model 7: MC-R predicted by VYesNo, WAT, OOPT (overall), IELTS (overall), age of starting to learn English and time spent living in the UK

How well did VYesNo, WAT, OOPT (overall) and IELTS (overall) scores, age of starting to learn English and time spent living in the UK predict MC-R scores, and which was the best predictor?

The total variance in MC-R scores explained by Model 7 as a whole was 52.1% ($R^2 = 0.521$), $F(6,96) = 17.384$, $p < .001$. Table 9.8 showed that the OOPT (overall) had a largest beta coefficient, 0.318 ($p = .001$), suggesting that this variable made the strongest unique contribution to explaining the variance in the MC-R scores when all other variance in the model was controlled for. IELTS (overall) ($\beta = 0.262$, $p = .004$) and the WAT ($\beta = 0.209$, $p = .021$) also made contributions significant at the .01 and .05 levels respectively, whereas VYesNo, age of starting to learn English and time spent living in the UK did not make any significant contribution to explaining the variance in MC-R.

[96] 46A and 52A completed their tests in the lab sessions but erroneously selected 'China' in response to 'where do you currently live?' and so automatically skipped the question about length of time spent living in the UK.

Table 9.8 *Model 7: Coefficients*[a]

Model		Unstandardized Coefficients		Stand. Coeffs.			95% Confidence interval for *B*		Correlations		
		B	*SE*	*Beta* (β)	*t*	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-48.004	15.372		-3.123	.002	-78.516	-17.491			
	VYesNo	.001	.001	.095	.993	.323	-.001	.003	.516	.101	.070
	WAT	.252	.107	.209	2.351	.021	.039	.464	.527	.233	.166
	OOPT (overall)	.289	.084	.318	3.444	.001	.122	.455	.616	.332	.243
	IELTS (overall)	6.442	2.169	.262	2.970	.004	2.136	10.748	.542	.290	.210
	Age start. Eng.	.357	.358	.072	.996	.322	-.354	1.067	.133	.101	.070
	Time in UK	.095	.143	.052	.664	.509	-.189	.379	.272	.068	.047

[a] Dependent variable: MC-R.

In other words, for every one unit increase in OOPT (overall), IELTS (overall) and the WAT (i.e., every mark gained or word associate recognised), an increase in MC-R of .318, .262 and .209 *SD* units respectively could be expected, controlling for the effect of the other predictors.

By squaring the part correlations, it was revealed that 5.9% of variance in MC-R was uniquely explained by the OOPT (overall), with any overlap or shared variance partialled out or removed, 4.4% by IELTS (overall), and 2.8% by the WAT.

In summary, the model, which included controls of VYesNo, WAT, OOPT (overall), IELTS (overall), age of starting to learn English and time spent living in the UK, explained 52.1% of the variance in MC-R scores. Of these six variables, the OOPT (overall) made the largest contribution (β = 0.318), significant at the .01 level, whilst IELTS (overall) and the WAT also made smaller contributions, significant at the .01 and .05 levels respectively.

9.4.2.2 Model 8: MC-P predicted by VYesNo, WAT, OOPT (overall), IELTS (overall), age of starting to learn English and time spent living in the UK

How well did VYesNo, WAT, OOPT (overall) and IELTS (overall) scores, age of starting to learn English and time spent living in the UK predict MC-P scores, and which was the best predictor?

The total variance in MC-P scores explained by Model 8 as a whole was 50.7% (R^2 = 0.507), $F(6,97) = 16.605$, $p < .001$. Table 9.9 shows that the WAT had a largest beta coefficient, 0.414 ($p < .001$), suggesting that this variable made the strongest unique contribution to explaining the variance in the MC-P scores when all other variance in the model was controlled for. The OOPT (overall) (β = 0.237, p = .013) and IELTS (overall) (β = 0.223, p = .014) also made statistically significant contributions at .05 level, whereas VYesNo, age of starting to learn English and time spent living in the UK did not make any discernible contribution to explaining the variance in MC-P.

Table 9.9 *Model 8: Coefficients*[a]

Model		Unstandardized Coefficients		Stand. Coeffs.			95% Confidence interval for *B*		Correlations		
		B	*SE*	*Beta* (β)	*t*	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-113.530	21.066		-5.389	.000	-155.339	-71.720			
	VYesNo	.000	.001	-.016	-.162	.871	-.003	.003	.449	-.016	-.012
	WAT	.676	.147	.414	4.607	.000	.385	.967	.613	.424	.329
	OOPT (overall)	.292	.115	.237	2.545	.013	.064	.521	.553	.250	.181
	IELTS (overall)	7.469	2.973	.223	2.512	.014	1.568	13.369	.478	.247	.179
	Age start. Eng.	.758	.491	.113	1.545	.126	-.216	1.732	.171	.155	.110
	Time in UK	.030	.196	.012	.155	.877	-.358	.419	.209	.016	.011

[a] Dependent variable: MC-P.

In other words, for every one unit increase in WAT, OOPT (overall), and IELTS (overall) (i.e., every word associate recognised or mark gained), an increase in MC-P of .414, .237 and .223 *SD* units respectively could be expected, controlling for the effect of the other predictors.

By squaring the part correlations, it was revealed that 10.8% of variance in MC-P was uniquely explained by the WAT, with any overlap or shared variance partialled out or removed, 3.3% by OOPT (overall), and 3.2% by IELTS (overall).

In summary, the model, which included controls of VYesNo, WAT, OOPT (overall), IELTS (overall), age of starting to learn English and time spent living in the UK, explained 50.7% of the variance in MC-P scores. Of these six variables, the WAT made the largest contribution (β = 0.414), significant at the .01 level, whilst the OOPT (overall) and IELTS (overall) made smaller contributions, significant at the .05 level.

9.4.2.3 Model 9: MC-R&P predicted by VYesNo, WAT, OOPT (overall), IELTS (overall), age of starting to learn English and time spent living in the UK

How well did VYesNo, WAT, OOPT (overall) and IELTS (overall) scores, age of starting to learn English and time spent living in the UK predict MC-R&P scores, and which was the best predictor?

The total variance in MC-R&P scores explained by Model 9 as a whole was 59.1% (R^2 = 0.591), $F(6,96) = 23.115$, $p < .001$. Table 9.10 shows that the WAT had a largest beta coefficient, 0.362 ($p < .001$), suggesting that this variable made the strongest unique contribution to explaining the variance in the MC-R&P scores when all other variance in the model is controlled for. The OOPT (overall) (β = 0.300, $p = .001$) and IELTS (overall) (β = 0.253, $p = .002$) also made statistically significant contributions at the .01 level, whereas VYesNo, age of starting to learn English and time spent living in the UK did not make any discernible contribution to explaining the variance in MC-R&P.

Table 9.10 *Model 9: Coefficients*[a]

Model	Unstandardized Coefficients	Stand. Coeffs.	*t*	Sig.	95% Confidence interval for *B*	Correlations

		B	*SE*	*Beta* (β)			Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	-74.136	14.867		-4.987	.000	-103.647	-44.625			
	VYesNo	.000	.001	.026	.294	.770	-.002	.002	.516	.030	.019
	WAT	.456	.104	.362	4.406	.000	.251	.662	.628	.410	.288
	OOPT (overall)	.285	.081	.300	3.515	.001	.124	.446	.632	.338	.229
	IELTS (overall)	6.517	2.098	.253	3.106	.002	2.353	10.682	.545	.302	.203
	Age start. Eng.	.385	.346	.075	1.112	.269	-.302	1.073	.142	.113	.073
	Time in UK	.087	.138	.045	.628	.531	-.188	.361	.265	.064	.041

[a] Dependent variable: MC-R&P.

In other words, for every one unit increase in WAT, OOPT (overall), and IELTS (overall) (i.e., every word associate recognised or mark gained), an increase in MC-R&P of .362, .300 and .253 *SD* units respectively could be expected, controlling for the effect of the other predictors.

By squaring the part correlations, it was revealed that 8.3% of variance in MC-R&P was uniquely explained by the WAT, with any overlap or shared variance partialled out or removed, 5.2% by OOPT (overall), and 4.1% by IELTS (overall).

In summary, the model, which included controls of VYesNo, WAT, OOPT (overall), IELTS (overall), age of starting to learn English and time spent living in the UK, explained 59.1% of the variance in MC-R&P scores. Of these six variables, the WAT made the largest contribution (β = 0.362), closely followed in size by the OOPT (overall) and to a lesser extent by IELTS (overall); all were significant at the .01 level.

9.4.2.4 Magnitude of predictive power: Hierarchical regression

The predictive values of the independent variables were explored further using hierarchical regressions aimed at investigating the magnitude of R^2 changes. For MC-R, with OOPT (overall) (best predictor) entered into the model at the first step and the other variables entered at the second step, the R^2 change was .141 (*F* change = 5.649, $p < .01$). With IELTS overall (second best predictor) entered into the model at the first step and the other variables entered at the second step, the R^2 change was .227 (*F* change = 9.099, $p < .01$). These data indicate that the rest of the variables provided an additional 14.1% of the criterion (DV) variance over and above the OOPT (overall) (best predictor), and an additional 22.7% over and above IELTS overall (second best predictor).

For MC-P, with the WAT (best predictor) entered into the model at the first step and the other variables at the second, the R^2 change was .130 (*F* change = 5.125, $p < .01$). With OOPT (overall) (second best predictor) entered into the model at the first step and the other variables entered at the second step, the R^2 change was .201 (*F* change = 7.888, $p < .01$). These data indicate that the rest of the variables provided an additional 13% of the criterion (DV) variance over and above the WAT (best predictor), and an additional 20.1% over and above OOPT (overall) (second best predictor).

For MC-R&P, with the WAT (best predictor) entered into the model at the first step and the other variables at the second, the R^2 change was .196 (F change = 9.201, $p < .01$) for MC-R&P. With OOPT (overall) (second best predictor) entered into the model at the first step and the other variables entered at the second step, the R^2 change was .192 (F change = 9.003, $p < .01$). These data indicate that the rest of the variables provide an additional 19.6% of the criterion (DV) variance over and above the WAT (best predictor), and an additional 19.2% over and above OOPT (overall) (second best predictor).

9.4.2.5 Summary

In summary, VYesNo, WAT, OOPT (overall), IELTS (overall), age of starting to learn English and time spent living in the UK were able to significantly predict 52.1%, 50.7% and 59.1% of the total variance in MC-R, MC-P and MC-R&P scores respectively. Whereas the OOPT (overall) was the best predictor of MC-R, the WAT was the best predictor of MC-P and MC-R&P.

9.5 Confirming the non-effect of 'test setting' on the data

When the possible effect of 'test setting' was controlled for, were the independent variables in the above analyses still able to predict significant amounts of variance in dependent variables?

As reported in the methodology chapter, while the test protocol was the same for all NNS participants, some of these participants (n = 35) completed tests individually during group sessions in a computer lab, while the remainder (n = 77) completed the tests at home. Although a comparison of means showed no MC test score differences for lab and home groups (section 4.6.4), this could be further confirmed by rerunning all the analyses reported above as hierarchical regression analyses with 'test setting' statistically controlled for (i.e., entered as a block 1 control before other independent variables were entered as block two variables).

The model summaries showed that after 'test setting' was entered as a block 1 variable, R^2 values ranged from 0.005 to 0.029 (explaining 0.5% to 2.9% of the total variance in DVs). In no case was the F change (from no model to the block 1 model) significant at the .05 level. After block 2 variables were entered, R^2 values ranged from 0.351 to 0.593 (explaining 35.1% to 59.3% of the total variance). In each case the F change was significant at the .01 level. These results provided further confirmation of the non-effect of 'test setting' in the present study. Regardless of whether participants completed the tests in lab sessions or at home, in all cases, the independent variables are still able to predict significant amounts of variance in the dependent variables.

9.6 MC-R and MC-P correlations at different L2 proficiency levels

RQ6: To what extent is the relationship between L2 receptive and productive metaphoric competence different at various L2 proficiency levels?

9.6.1 Data preparation

The final research question investigated was the extent of any change in correlation between MC-R and MC-P from lower to higher L2 proficiency levels and eventually, the NS level. The first step in answering this question was to arrange NNSs into different L2 proficiency groups. Groups were established according to the OOPT (overall) scores indexed as CEFR levels ranging from A2 to C2+. Since the OOPT (overall) scores had been directly collected by the researcher, whereas IELTS (overall) had been reported by participants, the former was preferable for this purpose. While the range A2 to C2+ suggested six CEFR proficiency levels, in order to ensure sufficient sample sizes for the correlational analyses and to allow for comparison with other studies, NNSs were parsed into three general proficiency ranges: 'low' (B1 or less), 'mid' (B2) and 'high' (C1 or above). Importantly, the labels 'low', 'mid' and 'high' were assigned to facilitate interpretation of results relative to the present sample, rather than as objective descriptors.

9.6.2 Results

Table 9.11 below shows that the MC-R and MC-P correlations were 'medium', 'large', 'small' and 'negligible' (Pallant, 2013, p. 139) for the 'low', 'mid', and 'high' NNS groups and the NS group respectively.

Table 9.11 *Correlations between MC-R and MC-P at Different L2 Proficiency Levels*

	NNSs			NSs
	'Low' group	'Mid' group	'High' group	
N	33	50	26	22
Correlation (95% CIs)	.39* (.06,.65)	.55 (.32,.72)	0.29 (-.12,.61)	0.00 (-.39,.40)

Note. Groups formed according to OOPT (overall) scores.

* Correlation is significant at the 0.05 level (2-tailed).

Only the correlation for the 'low' group was significant (at the .05 level), however, the fact that both lower and upper bound 95% confidence intervals were positive for the 'low' and 'mid' groups, but not the 'high' and 'NS' groups suggests that MC-R and MC-P appear to be correlated at lower L2 proficiency levels and not correlated at higher L2 proficiency levels and the NS level. The general downward trend can be seen in Figure 9.1.

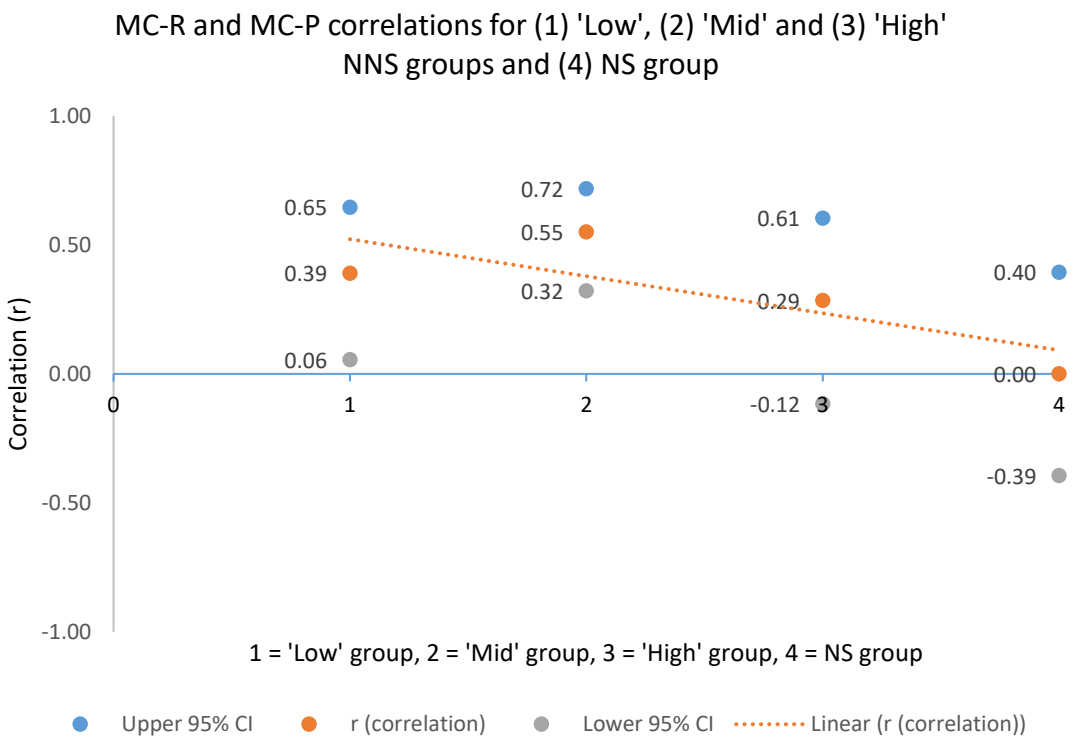


Figure 9.1 Scatterplot: MC-R and MC-P correlations for (1) 'Low', (2) 'Mid' and (3) 'High' NNS groups, and (4) NS group

The magnitude and significance of these correlation changes was assessed using a function of the VasserStats resource (Lowry, 2017). Table 9.12 below shows the greatest magnitudes of change from the 'Mid' to 'NS' group, significant at the .01 level (one-tailed). The next greatest correlation changes were from the 'Low' to NS and 'Mid' to 'High' groups, however these were not significant at the .05 level.

Table 9.12 *Magnitude and Significance of MC-R and MC-P Correlation Differences ('Low', 'Mid', and 'High' NNS Groups, and NS Group)*

		Sig.	
Groups[a]	*Z* value	One-tailed	Two-tailed
Low to Mid	-0.89	0.19	0.37
Low to High	0.43	0.33	0.67
Low to NS	1.41	0.08	0.16
Mid to High	1.28	0.10	0.20
Mid to NS	2.28	0.01	0.02
High to NS	0.94	0.17	0.35

[a] Formed according to OOPT (overall) scores.

In summary, at around the B2 level, MC-R and MC-P correlations appeared to decrease, with negligible correlations between these two sets of scores for higher L2ers and NSs. This trend is discussed in the next chapter.

9.7 Chapter summary

The focus of this chapter was on investigating:

a) the extent to which L2 vocabulary knowledge (size and depth), L2 proficiency (OOPT and IELTS), age of starting to learn English and time spent living in the UK predict L2 metaphoric competence test scores (RQ5)
b) the extent of the relationship between L2 receptive metaphoric competence and L2 productive metaphoric competence at various L2 proficiency levels (up to the NS level) (RQ6)

Regression 1 showed that in combined models, VYesNo and WAT were able to significantly predict 35%, 39.4% and 43.6% variance in MC-R, MC-P and MC-R&P scores respectively. Both VYesNo and WAT had significant, predictive power, and while the WAT was a better predictor in all modes, the two vocabulary tests always explained more MC-R, MC-P and MC-R&P variance in combination.

Regression 2 showed that in combined models, the OOPT and IELTs components were able to significantly predict 45.2%, 37.1% and 46.5% of the total variance in MC-R, MC-P and MC-R&P scores respectively. The OOPT Listening was the best predictor of MC-R, whereas the OOPT Use of English was the best predictor of MC-P and MC-R&P, however, the proficiency components always explained more MC-R, MC-P and MC-R&P variance in combination.

Regression 3 showed that combined in models VYesNo, WAT, OOPT (overall), IELTS (overall), age of starting to learn English and time spent living in the UK were able to significantly predict 52.1%, 50.7% and 59.1% of the total variance in MC-R, MC-P and MC-R&P scores respectively. The OOPT (overall) was the best predictor of MC-R, while the WAT was the best predictor of MC-P and MC-R&P. All of these measures, however, always explained more MC-R, MC-P and MC-R&P variance in combination.

A rerun of all regression analyses but with 'test setting' entered hierarchically as a block 1 variable showed that whether NNSs took the tests in a 'lab' setting or at 'home' had negligible effects on MC-R, MC-P and MC-R&P scores.

Finally, a correlation analysis showed a medium-to-large strength relationship between MC-R and MC-P for 'low' (B1 or less) and 'mid' (B2) level NNSs, but negligible correlations between these two variables for 'high' (C1 or above) NNSs and NSs. These findings are discussed in the following chapter.

Chapter 10: Discussion of Analysis 3

10.1 Introduction

This chapter will discuss the key findings of Analysis 3: Relationships between L2 metaphoric competence, vocabulary knowledge, general proficiency, age of starting to learn English and time spent living in the UK in relation to previous metaphoric competence and other research. In keeping with the previous discussions, this chapter is structured into two main parts corresponding to the fifth and sixth research questions, with subsections on emerging themes.

10.2 RQ5: To what extent can L2 vocabulary knowledge (size and depth), L2 proficiency (Oxford Online Placement Test and IELTS), age of starting to learn English and time spent living in the UK predict L2 metaphoric competence test scores?

10.2.1 L2 metaphoric competence predicted by L2 vocabulary size and depth

Regression 1 showed that in combination, VYesNo and WAT were able to significantly predict 35%, 39.4% and 43.6% of the total variance in MC-R, MC-P and MC-R&P scores respectively (models 1-3). Both VYesNo and WAT had significant predictive power (at the .01 level), and offered more explanatory power in combination than any one in isolation. However, the WAT proved superior for all modes. Out of all the regression 1 analyses, the best individual prediction was the WAT as a predictor of MC-P scores, to which VYesNo added very little additional explanatory power.

In terms of correlation, VYesNo and WAT were quite strongly related to MC-R, MC-P, and MC-R&P. This is somewhat unsurprising given the generally strong correlations between vocabulary knowledge measures and L2 MC observed in past research on a comparable east Asian sample (Azuma, 2005). However, when it comes to regression, the absence of any analyses exploring the ability of L2 vocabulary knowledge measures (or any other variable for that matter) to predict L2 MC scores problematises interpretation of the present results. As a way around this problem, the behaviour of L2 vocabulary knowledge measures as predictors of more general (i.e., non-MC) areas of L2 proficiency (for which several findings exist) were considered.

For instance, Qian's (2002) finding that the VLT (vocabulary size), DVK (vocabulary depth) and TOEFL-Vocabulary Item Measure in combination led to better predictions of L2 reading comprehension than any one alone mirrors the present study's discovery that VYesNo and WAT explained more L2 MC variance in combination. In this respect, Qian's study and this

one highlight the conceptual overlap between vocabulary size and depth as part of language competence, since a combined increase in both (rather than size or depth individually) are likely to have the most benefit on language competence. Assessing the magnitude of explanatory power in the vocabulary measures used in regression 1 is also problematised due to the lack of comparable research. However, Schmitt's (2000, p. 4) comment that 42.6 per cent of variance in class grade scores explained by knowing L2 form-meaning word links (i.e., vocabulary size) is "a very great deal", seems to suggest that the amounts of variance explained in models 1-3 (and indeed further models) is quite substantial, particularly given all the other possible things, in addition to L2 vocabulary size and depth (as measured), that might explain variance in the L2 MC scores observed, but that were not measured in the present study (e.g., cognitive style, world knowledge, willingness to use the L2, creativity).

Despite their joint contribution, vocabulary depth construed as lexical organisation seems to have had some sort of an 'edge' over vocabulary size, particularly when predicting productive metaphoric competence scores. This finding is also born out in past research. For instance, in both the present study and Azuma (2005), receptive, productive and combined receptive and productive MC had slightly stronger correlations with vocabulary depth than vocabulary size,[97] albeit that Azuma's correlations were based on a different vocabulary depth measure to the one used in the present study. Remembering that MC-P tested form-recall, whereas MC-R predominantly tested form-recognition, the discovery that the WAT had a clearly stronger relationship with MC-P aligns with and extends research in this area (e.g., Greidanus et al., 2004) which confirm that form-recall, the most difficult aspect of vocabulary knowledge to acquire (Laufer & Goldstein, 2004), is more strongly related to lexical organisation than any other aspect of vocabulary knowledge is.

Taken together, these points suggest that *both* vocabulary size and depth (as measured) were important for L2 metaphoric competence (as measured), and both should, and to a certain extent will (Schmitt, 2014), be developed in tandem. However, increasing one's L2 lexical network (i.e., vocabulary depth) rather than learning to recognise new forms (vocabulary size) is likely to have the most positive impact on L2 metaphoric competence, particularly producing metaphor.

10.2.2 L2 metaphoric competence predicted by L2 proficiency components

Regression 2 showed that in combination, the OOPT and IELTs components were able to

[97] In the present study, correlations (all significant at the .01 level) were: r = .52 (MC-R and VYesNo); r = .45 (MC-P and VYesNo); r = .52 (MC-R&P and VYesNo); r = .53 (MC-R and WAT); r = .61 (MC-P and WAT); r = .63 (MC-R&P and WAT).

significantly predict 45.2%, 37.1% and 46.5% of the total variance in MC-R, MC-P and MC-R&P scores respectively (models 4-6). Combined models involving all variables always had more explanatory power than any individual variable in isolation. MC-R was most strongly predicted by the OOPT Listening, but also by OOPT Use of English (both had beta values significant at the .01 level). MC-P was best predicted by OOPT Use of English (beta significant at the .01 level), but also arguably by OOPT Listening and IELTS Listening (although betas were significant at the .1 level only). MC-R&P was best predicted by OOPT Use of English, but also by OOPT Listening (betas significant at the .01 level) and somewhat less so by IELTS Listening (significant at the .1 level only). Of all the regression 2 analyses, the best individual prediction, to which additional variables added the least additional explanatory power, was the OOPT Use of English as a predictor of MC-P.

Past studies have found a strong correlation between L2 receptive MC and L2 reading in L1 Chinese learners of English (Zhao et al., 2014), and a very strong between L2 receptive and productive MC and the 2001 version of the OOPT in L1 Persian learners of English (Aleshtar & Dowlatabadi, 2014). However, these studies have limited application to the present findings, since neither investigated the correlations of respective proficiency strands (reading, writing, speaking, listening), or of different L2 proficiency tests (e.g., OOPT and IELTS) as in regression 2. To the extent that the NNS sample are generalisable, the present study's findings have the following implication: A student who improves her ability to answer OOPT Listening questions is likely to see the greatest gains in MC-R scores, whereas an improvement in OOPT Use of English scores should yield comparatively higher MC-P (than MC-R) scores. What can account for this?

One possibility concerns the type of questions featured in OOPT Listening and OOPT Use of English tests. All three tasks in OOPT Listening exclusively measure form and meaning recognition skills through the multiple-choice format. The OOPT Use of English section on the other hand, involves both multiple-choice *and* unaided sentence completion/gap-fill tasks; in other words, both form- and meaning-recognition and recall. If these question formats are paralleled with those of the MC Test Battery, it becomes clear that the OOPT Listening tests skills more akin to those engaged by MC-R, which predominantly used multiple-choice, form-recognition questions, whereas MC-P test skills more akin to those tested by the OOPT Use of English, since it exclusively used limited production and form-recall. To the extent that this connection accounts for the observed predictions in regression 2, it highlights the importance of task type when considering possible explanations for an observed relationship between predictor and criterion variables.

Of the IELTS components, only Listening predicted a small amount of variance in MC-P and MC-R&P, however the beta value was significant at the .1 level only. This slightly confuses the above account because one might well ask how IELTS Listening, a test that measures a

receptive aspect of L2 proficiency, predict (albeit a small amount of) variance in productive but not receptive metaphoric competence scores? In addressing this question, it is important to again consider the task types involved in IELTS Listening: multiple-choice, matching, plan/map/diagram labelling, form/note/table/flow-chart summary completion, and sentence completion. As set out in the literature review, a number of IELTS Listening tasks (e.g., sentence completion (section 2.4.2.4) required test takers to both identify and supply the correct form when listening to a dialogue, penalising poor spelling and grammar. Such tasks are likely to have placed more of a cognitive burden on test takers than multiple-choice recognition (where even a blind guess has a 25% chance of being correct) and in this respect, IELTS Listening bears more resemblance to the MC-P recall tasks than the MC-R multiple-choice, acceptability rating and explain the meaning questions. In this view, the question is then why IELTS Listening, rather than Speaking, Reading, or Writing, predicted some variance in MC-P scores. This is particularly puzzling given the clear engagement of productive skills in IELTS Writing and Speaking.

A speculative answer to this question again concerns the type of task involved. IELTS Writing and Speaking both test 'free' (and in the case of the later), 'dynamic' and 'interactive' L2 production. In these tests, test takers ability to use cohesion, rhetorical organisation, and (for speaking) their handling of the trade-off between complexity, accuracy, fluency and lexical richness (Skehan, 2009) are critical determiners of the score they achieve. By comparison, both MC-R and MC-P required short answers, with no time pressure. Thus, neither tapped into the kind of 'online' processing required for IELTS Speaking, nor the discourse organisation skills required for IELTS Writing. In that sense they are more akin to IELTS Reading and Listening, which require shorter answers. So why was IELTS Reading not a good predictor of either MCR or MC-P? While the slightly higher correlation between MC-R and IELTS Reading than MC-P and IELTS Reading is somewhat intuitive given the discussion above, a more comprehensive answer to this question is the task of further research.

10.2.3 L2 metaphoric competence predicted by L2 vocabulary knowledge vs by L2 general proficiency (overall)

Regression 3 showed that in combination, VYesNo, WAT, OOPT (overall), IELTS (overall), age of starting to learn English and time spent living in the UK were able to significantly predict 52.1%, 50.7% and 59.1% of the total variance in MC-R, MC-P and MC-R&P scores respectively (models 7-9). MC-R was most strongly predicted by the OOPT (overall), but also by IELTS (overall) and the WAT. MC-P was best predicted by the WAT, but also by OOPT (overall) and IELTS (overall). MC-R&P was best predicted by the WAT, but also by the OOPT (overall) and IELTS (overall). Thus, for these participants, over half of the variance in L2 metaphoric competence (as measured) was explained by L2 vocabulary and proficiency. This suggests that vocabulary knowledge and

general proficiency are hugely important part of Low (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skills and (sub)competences, when elicited and measured in the written mode. The amount of variance (approximately 40-50%) left unexplained by these models is highly intriguing, and may point to the proportion of L2 metaphoric competence that is non-linguistic (e.g., conceptual, strategic).

Of all the regression 3 analyses, the best individual prediction, to which additional variables added the least additional explanatory power, was the WAT as a predictor of MC-P. Curiously, although the WAT had the highest beta value for MC-R&P, when variables were entered hierarchically, the OOPT Use of English (second best predictor) explained more variance than WAT (with remaining variables adding less additional explanatory power) than the other way around.

Although a general predictive relationship between L2 proficiency and MC-R, and L2 vocabulary knowledge and MC-P and MC-R&P might be concluded from these results, it is important to emphasise the particular explanatory power of the OOPT (overall) and WAT. While IELTS was also a good predictor of MC-R, one certainly cannot substitute the vocabulary size measure (VYesNo) and expect the same predictive effect for MC-P. Past research would also suggest that not all measures of L2 vocabulary depth would correlate equally strongly with L2 metaphoric competence (Azuma, 2005). That the OOPT (overall) and IELTS (overall) had stronger relationships with MC-R than MC-P is unsurprising, given previous research (Aleshtar & Dowlatabadi, 2014; Zhao et al., 2014). Likewise, Azuma's (2005) observation that a vocabulary depth measure was the strongest correlate of productive metaphoric competence arguably also aligns with the present study's finding that the WAT best predicted MC-P and MC-R&P. However, this comparison is imperfect, since Azuma did not also explore correlations between MC and L2 proficiency measures, and since correlation (unlike regression) does not produce estimates with variance shared between predictors controlled for.

MC-R&P was an equally weighted composite of MC-R and MC-P. Since MC-R and MC-P had different best predictors in regression 3, it was interesting to observe that the WAT rather than the OOPT (overall) was the best predictor of MC-R&P, when the predictive effects of other variables were controlled for. This is highly useful, because it suggests that seeking to achieve a greater connectedness between pre-existing representations in one's lexical network (Schmitt, 2014) may be the most worthy endeavour for improving L2 receptive and productive metaphoric competence, seemingly more so that increasing one's general proficiency. However, interpreting this finding is complicated by the fact that although the WAT had the highest beta value, the hierarchical regression showed that remaining variables explained slightly more variance above and beyond the WAT, than above and beyond the OOPT (overall)! This finding goes against the hierarchical regressions in 1 and 2, for which the remaining variables always

explain lower amounts of additional variance above and beyond the best predictor rather than the second best predictor. These patterns are perplexing, and require further investigation.

A comparison of the regression 2 and regression 3 models shows that slightly more variance in MC-R than MC-P was explained. Regression 1, on the other hand, had models explaining more variance in MC-P than MC-R. This anomaly can be explained by something that is common to regressions 2 and 3 but not regression 1, namely L2 proficiency measures. Seemingly, when only L2 vocabulary size and depth scores are required to explain L2 MC, more variance can be predicted in MC-P (than MC-R) scores. If both L2 vocabulary knowledge and general proficiency knowledge scores do the work, then it is possible to explain more variance in MC-R than MC-P.

In some ways, it was surprising that L2 vocabulary size, which in combination with the WAT predicted variance in metaphoric competence variables in regression 1, did not have any predictive power in combination with L2 proficiency measures in regressions 2 and 3. Since vocabulary size is so fundamental to L2 proficiency (Laufer & Goldstein, 2004); Schmitt (2000), this is probably explained by the fact that VYesNo shares a lot of variance with the OOPT and IELTS scores, and was thus controlled for in regression 3, yielding a low, non-significant beta value. Moreover, the fact that the OOPT (overall) and IELTS (overall) measure knowledge of language in grammatical and pragmatic context means the skills they test are more akin to those engaged by MC-R tasks, whereas VYesNo presents word forms in a decontextualized manner. This implies a large amount of conceptual overlap between the OOPT (overall), IELTS (overall) and MC-R in terms of skills tested.

10.2.4 Possible reasons why age of starting to learn English and time spent living in the UK did not predict L2 metaphoric competence

Another major finding from regression 3 was that age of starting to learn English and time spent in the UK had no detectable ability to predict MC-R, MC-P or MC-R&P scores. Concerning age of starting to learn English, it is somewhat unsurprising that this had no impact on L2 MC, for one because these data reveal nothing about the amount and type of learning NNSs actually engaged in, and second because participants may have had differing interpretations of what 'starting' to learning English actually meant. The non-effect of time spent living in the UK can be interpreted in at least two ways. The first approach would be to conclude from these data, and the low, non-significant correlation between time spent living in the UK and MC-R, MC-P, and MC-R&P that up to two years living in the UK (for these NNSs) was not enough to impact on their metaphoric competence. This finding matches Azuma's (2005) observation that up to one year abroad had no impact on five L1 Japanese participants' metaphoric(al) competence. Although Azuma's

sample is too small to draw hard and fast conclusions from, the same cannot be said for the present study, for which 107 participants' data were entered into the analysis. In the present study, three participants, who had lived in the UK for 3 years were deleted from the data as extreme cases, however, an examination of their MC-R and MC-P scores shows that even they did not appear to stand out from the rest.

To the extent that these findings speak to a non-effect of living in the UK, one might argue that the immersion setting does not, in and of itself, equate to unconscious engagement with metaphor, which must be noticed and purposefully nurtured by the learner. Even if learners did consciously engage with metaphor during their time in the UK, their efforts may have been hampered by a lack of autonomous skills to nurture understanding, dealing with gaps in knowledge, and producing metaphor appropriately (Littlemore & Low, 2006a). In addition, without instruction, L2 metaphoric competence may take several years to develop noticeably, even in an immersion setting. In this case, neither the participants in the present study nor Azuma's would have lived in an L2 immersion setting long enough to experience a real change in metaphoric competence.

A second approach to interpreting these findings would be to conclude that time spent in the UK had no effect on NNSs' L2 MC if measured offline and in the written mode, but it may have brought gains for L2 MC conceptualised as a fluid, online, interactive competence in the spoken mode (Littlemore, 2001). This is somewhat intuitive, given the positive effect that time spent living in the UK has on the acquisition of more diverse lexis and nativelike word combinations (Foster & Tavakoli, 2009), as well as better sensitivity to native- non-nativelike utterances for learners with higher phonological short-term memory (Foster et al., 2014) (section 2.3.3). Whether or not the present study's NNSs had higher levels of fluid, online, spoken L2 MC, undetected by the MC Test Battery, is unknown. Further research is needed, not only to explore whether the way in which L2 MC is conceptualised has an impact on whether L2 immersion gains are observed, but also to determine more generally the extent of overlap between written, offline L2 MC and spoken, online L2 MC within and between different NNSs. Despite some (flawed) cross-sectional research on L2 metaphoric competence at different proficiency levels (section 2.2.3.2), and studies exploring the short-term intervention gains for specific metaphor forms (e.g., Boers et al., 2014), there does not appear to have been any longitudinal investigation into the development of L2 MC in adults over a period of several years, either in a foreign or second language context.

10.3 RQ6: To what extent is the relationship between L2 receptive metaphoric competence and L2 productive metaphoric competence different at various L2 proficiency levels?

In order to investigate the final research question, NNSs were parsed into three general proficiency ranges: 'low' (B1 or less), 'mid' (B2) and 'high' (C1 or above). Although grouping NNSs in this way may be criticised for being a crude approach that loses the benefits of the original scaled measurement (Plonsky & Oswald, 2017), since the three NNS groups reflected quite a range of L2 proficiency, it was deemed sufficient for detecting MC-R and MC-P correlation differences. In its use of different L2 proficiency groups, the present study effectively resembled approaches taken by researchers seeking to decipher whether the direction of correlation strength change between vocabulary size and depth as L2 proficiency increases (Henriksen, 2008; Noro, 2002; Nurweni & Read, 1999; Schmitt & Meara, 1997).

The findings showed that MC-R and MC-P had a medium-to-large strength relationship for the 'low' and 'mid' NNS groups, evidenced by the positive lower and upper bound 95% confidence intervals, and negligible correlations for the 'high' NNS group and 'NS group, evidenced by the fact that confidence intervals passed through zero. This suggests that for these participants, the correlation between MC-R and MC-P has a general downward trend from lower to higher (and eventually NS) proficiency.

Although past research has shown that correlations between receptive and productive metaphoric competence for intermediate level participants had been mixed, showing both medium-to-large strength correlations (Azuma, 2005) and negligible (and non-significant)-to-small correlations (Littlemore, 2001), there is no comparable metaphoric competence study against which to compare this trend. To circumvent this problem, the vocabulary knowledge literature offers some comparable findings. For instance, decreases in correlations between the receptive and productive WAT, and between the receptive VLT and productive WAT were observed In L1 Danish learners from lower to higher high school grades (i.e., from lower to higher L2 proficiency) (Henriksen, 2008). In addition, that fact that meaning and form recognition gains have been shown to be accompanied by much smaller form recall gains suggests a closer association between receptive (i.e., recognition) than productive (i.e., recall) skills and L2 development (Laufer & Goldstein, 2004; Schmitt, 2010). To the extent that L2 receptive and productive MC behave like L2 receptive and productive vocabulary knowledge, this would predict a general decrease in the correlations from lower to higher L2 proficiency levels, which is which is what was observed.

These results indicate that at lower levels, recognising and recalling the metaphors that one knows goes hand in hand, whereas at a certain point (around the B2 level, it seems), the

relationship between receptive and productive MC starts to wane, eventually becoming negligible. The fact that receptive MC was unrelated to productive MC for the highest L2 proficiency NNS group, and for the NSs, could be because individual, creative and stylistic freedom is more pronounced at higher L2 (and native) proficiency. Although this explanation is speculative, and requires further enquiry, to the extent that L2 receptive and productive metaphoric competence behave like L2 receptive and productive vocabulary knowledge, research would suggest that engaging in language learning tasks designed to foster productive knowledge is likely to bring about greater gains for both L2 receptive and productive metaphoric competence, than tasks aimed at fostering receptive knowledge (Webb, 2005).

In addition to trying to validate the patterns witnessed in these data, further research could viably seek to establish at what point learners really gain control over the metaphors they produce and whether, for instance, they become more selective as productive knowledge increases.

10.4 Chapter summary

In this chapter, the key findings of **Analysis 3: Relationships between L2 metaphoric competence, vocabulary knowledge, general proficiency, age of starting to learn English and time spent living in the UK** were discussed in relation to previous metaphoric competence and other research.

Several aspects of the fifth research question were addressed. Concerning the predictive power of vocabulary size and depth, despite no MC study allowing for direct comparison, the discovery that VYesNo and WAT explained more L2 MC variance in combination than on their own aligned with past research into these constructs as predictors of L2 reading comprehension (Qian, 2002). The present study, however, extends these findings to metaphoric (rather than merely 'language') competence, and the productive mode. Given all the other possible things that might account for variance in L2 MC scores, the vocabulary size and depth measures had quite a substantial amount of explanatory power.

While these results emphasised the conceptual relatedness of VYesNo and WAT as predictors of L2 MC (Schmitt, 2000), vocabulary depth construed as lexical organisation (section 2.3.4) seemed to have an 'edge' over vocabulary size, particularly when predicting productive metaphoric competence scores. This finding generally aligns with past MC research (Azuma, 2005) and vocabulary knowledge research (Greidanus et al., 2004). This suggested that while increasing both vocabulary size and depth would be likely to boost L2 MC, it is increases in the L2 lexical network (i.e., vocabulary depth) rather than learning to recognise new forms (vocabulary size) that are likely to have the most positive impact on L2 MC, particularly producing metaphor.

Concerning the predictive power of various L2 proficiency components in relation to one another. The finding that OOPT Listening best predicted receptive metaphoric competence, whereas OOPT Use of English best predicted productive (and combined receptive and productive) MC was explained in terms of task type, namely that multiple-choice recognition questions used in the OOPT Listening making the skills tested more akin to those engaged by the MC-R, whereas the mixture of multiple-choice recognition and gap-fill recall tasks in OOPT Use of English meant this test was more similar to MC-P. The ability of IELTS Listening, rather than any other IELTS component, to explain scores in MC-P and MC-R&P was interpreted as reflecting the fact that IELTS Listening largely requires test takers to identify and supply the correct form when listening to a dialogue (making the skills tested more akin to those engaged by MC-P than MC-R), and that the other, more obviously productive components (IELTS Speaking and Writing) did not have predictive power since they elicit skills not measured by the MC Test Battery (e.g., lexical cohesion, rhetorical organisation, fluency, lexical richness).

Another aspect of research question five concerned the comparative predictive power of the L2 vocabulary knowledge and L2 proficiency measures. The fact that models 7, 8 and 9 explained over half of the variance in L2 metaphoric competence (as measured) suggests that vocabulary knowledge and general proficiency are integral to Low (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skills and (sub)competences, when elicited and measured in the written mode. In addition, the magnitude of unexplained variance is important for future research considering the proportion of L2 metaphoric competence that is non-linguistic (e.g., conceptual, strategic). However, to properly understand these findings, one must go further and look at the explanatory power of specific measures.

The finding that the OOPT (overall) better predicted receptive MC, whereas the WAT better predicted productive MC was expected given past research on the relationships between receptive MC and L2 proficiency (Aleshtar & Dowlatabadi, 2014; Zhao et al., 2014), and productive MC and L2 vocabulary depth (Azuma, 2005). The fact that VYesNo had no predictive power when included in models alongside L2 proficiency measures (but did when included with vocabulary depth alone) was interpreted as indicating high levels of shared variance and conceptual relatedness between these two aspects of language competence. More specifically, L2 proficiency measures explain everything that vocabulary size does, and more. The ambiguity over whether the WAT or OOPT (overall) was the best predictor of MC-R&P was noted as requiring further research, but from the present findings, it seems there is again a distinct advantage for L2 metaphoric competence in seeking to strengthen the connections in one's pre-existing lexical network.

Age of starting to learn English and time spent living in the UK were found to have no power to predict L2 MC. The non-effect of age of starting to learn English was attributed to the

fact that this variable revealed nothing about the amount and type of L2 learning NNSs actually engaged in, and because the meaning of 'starting' may have been interpreted different from NNS-to-NNS. Two interpretation were offered to explain the non-effect of time spent living in the UK. First, in line with Azuma (2005), it is possible that up to two years living in the UK (for the present study's NNSs) was not enough to impact on their L2 MC, and that that the immersion setting, in and of itself, does not necessitate unconscious engagement with metaphor, which needs to be noticed and is likely to require focused instruction (Littlemore & Low, 2006a). Alternatively (but not mutually exclusively), the non-effect of time spent living in the UK may indicate that while no L2 MC gains were observed in the written, offline mode (i.e., via the MC Test Battery), there may have been gains in the online, spoken mode, however these would have been undetected in the present study. It was concluded that further, particularly longitudinal, research is needed to determine the effect of time spent in an L2 immersion setting on L2 MC, and the extent to which offline, written MC relates to online, spoken MC.

Finally, concerning the sixth research question, a comparison of the correlations between MC-R and MC-P for 'low', 'mid' and 'high' proficiency NNS groups and the NS group showed a general downward correlational trend. While no comparable L2 MC study had investigated this question, the findings found some parallel with correlational trends between receptive and productive vocabulary measures (Henriksen, 2008), and the observation that receptive (i.e., recognition), rather than productive (i.e., recall) skills are more closely bound to L2 proficiency (Laufer & Goldstein, 2004; Schmitt, 2010). The decreasing trend, from medium-to-large strength correlations in the 'low' (CEFR B1 or less) and 'mid' (CEFR B2) NNS groups, to negligible correlations for the 'high' (CEFR C1 and above) NNS group and NS group, was tentatively interpreted as pointing to the more influence of other factors (e.g., individual, creative and stylistic freedom) being more pronounced at higher L2 (and native) proficiency, thus leading to greater variation between individuals in terms of their productive MC.

This concludes the analyses and discussion chapters. In the final chapter, general conclusions from the study are drawn, limitations, future research directions and tentative teaching implications are noted, and the study's contribution to the existing research literature stated.

Chapter 11: Conclusion

11.1 Summary of the study

This thesis has presented the findings of a linguistic-based investigation into the L2 metaphoric competence of L1 Chinese NNSs of English. Specifically, the thesis explored the extent to which L2 metaphoric competence can be reliably measured, its subcomponents, and its relationship with L2 vocabulary knowledge and general proficiency, as well as age of starting to learn English and time spent living in the UK. The participants were 112 NNSs of English (L1 Chinese) enrolled or already engaged in undergraduate or postgraduate study at UK Universities and 31 (British) English NSs, either engaged in postgraduate study, in employment or recently retired.

The study was divided into three analyses. The first sought to understand the extent to which a battery of L2 metaphoric competence tests could reliably elicit and measure Low (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skill and (sub)competences (RQ1) and NNS and NS group differences between MC test scores (RQ2). The second analysis investigated the extent to which factors underlie the L2 MC, vocabulary knowledge and general proficiency scores and the nature of these (RQ3), whether the same factors could be observed in combined NNS and NS data and how NNS (L1 Chinese) and NS (L1 English) factors scores differed (RQ4). The third analysis explored the extent to which L2 vocabulary knowledge (size and depth), L2 proficiency (OOPT and IELTS), age of starting to learn English and time spent living in the UK could predict L2 MC test scores (RQ5), and the extent of the relationship between L2 receptive and L2 productive metaphoric competence at different L2 proficiency (up to NS) levels (RQ6).

The MC Test Battery was developed and refined via pre-pilot and pilot studies involving similar NNS and NS participants. Vocabulary size was measured using the VYesNo test (Meara & Miralpeix, 2015) vocabulary depth was measured using the Word Associates Test (1993), and L2 proficiency was measured using the OOPT and (reported) IELTS scores. In order to minimise NNSs' anxiety when taking the MC tests (Azuma, 2005) and in order to recruit a sufficiently large sample of L2 learners for regression analysis (Plonsky, 2013), NNSs were offered the choice of completing tests at pre-arranged 'lab' sessions, or at 'home' in their own time. Although 35 NNSs completed the tests in 'lab' sessions while 77 completed them at home, the different test settings had no observable effect on the data, demonstrated by the absence of any statistically significant differences ($p < .01$) between 'lab' and 'home' group test scores, and the non-effect of 'test setting' in regression analyses. All NSs completed the test at 'home'. After completing tests, NNSs received £5 cash or a £5 Amazon voucher on the spot. After all data had been collected, NNSs were emailed feedback and invited to a session to discuss the tests and further

ways to practice and improve their L2 metaphoric competence. NSs were emailed feedback upon request.

11.2 Summary of the findings

Analysis 1: The development and reliability of the MC Test Battery, and descriptive statistics, showed that the extent to which L2 MC could be reliably elicited and measured was complicated by problems concerning mixed levels of (particularly instrument) reliability, problems eliciting metaphor, likely test taker anxiety, forming an orderly approach to data cleaning, high levels of NS variation for tasks involving acceptability of Vehicle extensions (cf. Low, 1988), and the applicability of response data obtained from receptive multiple-choice questions to the real-world and to pedagogy (RQ1). Some metaphor-related skills or (sub)competences it seems (e.g., idiom extension), are more reliably measured than others. Several areas where the approach taken aided reliable measurement of Low (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skill and (sub)competences were noted. These included the extensive deletion of problematic items, tests and participants order to yield a truer representation of the constructs targeted (Littlemore, 2001), and the refinement of task instructions, examples, explanations and ordering of tests to facilitate and motivate test takers (points which researchers using the MC Test Battery in future research might take on board).

Descriptive statistics showed that several of the MC tests expected to be particularly challenging for the NNSs (e.g., Test 8-Idiom Extension-R and –P), or where large NNS and NS differences were expected to be observed (e.g., Test 1-Phrasal Verbs-R and –P) were found to be so (RQ2). A close analysis of the response data for Test 8-Idiom Extension–P showed that although some idiom extensions elicit more avoidance than others, the NNSs were generally quite capable of engaging in this task (cf. Littlemore & Low, 2006a). The fact that NSs showed ceiling effects for several tests (Test 1-Phrasal Verbs-R and –P, Test 5-Topic Transition-R, and Test 7-Feelings) point to areas of L1 MC which might be similar for all adult L1 speakers, and thus considered more prototypically part of Basic Language Cognition (Hulstijn, 2011, 2012), although this is a tentative suggestion requiring substantial verification.

Analysis 2: Metaphoric and other (sub)competences uncovered reported an EFA of L2 MC, vocabulary knowledge, and L2 proficiency test scores. This analysis aimed to address several shortcomings of past EFAs involving L1 MC variables (e.g., inappropriate factor extraction technique, reporting measures of model adequacy, consistent use of and empirical basis for qualitative descriptors 'high', 'strong', 'stable'). Despite different factor retention criteria suggesting disparate numbers of factors to retain, the EFA of the NNS data showed that 42% of variance in L2 MC, vocabulary knowledge and general proficiency test scores could be explained by a six-factor solution. Using a principled method, four of the factors were identified as L2 MC

(sub)competences: **English Grammatical Metaphoric Competence**, **English Illocutionary Metaphor Production**, **English Metaphor Language Play**, and **English Topic/Vehicle Acceptability**; and two of the factors appeared to be distinct and largely non-MC constructs: **English Vocabulary Size**, and **English General Comprehension** (RQ3). The mixed homogeneity and heterogeneity of factors with regard to Low (1988) and Littlemore and Low's (2006a) framework components suggested further reconsideration and potential revision of the authors' frameworks.

The EFA of the NNS+NS data showed that the four MC factors (albeit with variations in loading variables) re-emerged, and that NNSs (L1 Chinese) and NSs (L1 English) differed in their overall MC (i.e., for all factors) but only in terms of their **English Grammatical Metaphoric Competence**, when individual factors were considered (RQ4). These results suggested that **English Grammatical Metaphoric Competence** was the hardest to acquire aspect of L2 MC, both for the present study's NNSs, and also (it is likely, but needs verifying) for learners from other L1s (Kellerman, 1983; Liao & Fukuya, 2004). A tentative case study of why NNSs and NSs differed for **English Grammatical Metaphoric Competence** showed patterns in the *incorrect* particles that NNSs selected and produced, but that form frequency was an inadequate proxy for determining phrasal verb item difficulty. Some examples of NNS and NS responses for **English Illocutionary Metaphor Production**, **English Metaphor Language Play**, and **English Topic/Vehicle Acceptability** test items were given to highlight qualitative differences (but not deficiencies) between NNSs and NSs for these statistically 'equivalent' (sub)competences. Although not part of the scoring, many of these differences seemed to concern phraseological proficiency (Philip, 2010).

Analysis 3: Relationships between L2 metaphoric competence, vocabulary knowledge, general proficiency, age of starting to learn English and time spent living in the UK involved three regression analyses focusing on different aspects of RQ5. Regression 1 found that the VYesNo and WAT explained more L2 MC variance in combination than on their own, the WAT had slight predictive 'edge' for all modes, particularly productive MC, a finding that is generally parallel in past research (Azuma, 2005; Greidanus et al., 2004). To the extent that these findings generalise, this suggested that increasing links between the representations in one's lexical network, rather than acquiring new form-meaning links, would likely be more beneficial for developing L2 MC, although the two are somewhat inseparable (Schmitt, 2014). Comparing OOPT and IELTS proficiency components, OOPT Listening best predicted receptive metaphoric competence, whereas OOPT Use of English best predicted productive (and combined receptive and productive) MC. This was explained by the fact that the OOPT Listening exclusively involved multiple-choice recognition tasks, whereas OOPT Use of English involved multiple-choice recognition and gap-fill recall tasks, meaning the skills it tests are more akin to those engaged

by MC-R and MC-P respectively.

Comparing the predictive power of all vocabulary, proficiency, age and immersion measures, the OOPT (overall) was the best predictor of L2 receptive MC (Aleshtar & Dowlatabadi, 2014; Zhao et al., 2014), whereas the WAT was the best predictor of L2 productive MC (Azuma, 2005). The fact that models 7, 8 and 9 explained over half of the variance in L2 metaphoric competence (as measured) was highly revealing, because it showed that Low (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skills and (sub)competences (as measured) are largely synonymous with these measures. The proportion of L2 metaphoric competence that was not explained by these models is likely to be non-linguistic (e.g., conceptual, strategic), however uncovering these is the task of further research. The ability of the VYesNo to predict L2 MC in regression 1 but not in regression 3 was interpreted as showing that the OOPT (overall) and IELTS (overall) explain everything that vocabulary size does, and more, thus cancelling its predictive power out in combined models. The ambiguity over whether the WAT or OOPT (overall) was the best predictor of MC-R&P was noted as requiring further research, however these findings suggest that it is specifically increasing the strength of connections in one's pre-existing lexical network (rather than general L2 proficiency) that is likely to bring the highest gains for L2 metaphoric competence. Interestingly, age of starting to learn English and time spent living in the UK had no ability to predict L2 MC, (the latter) explained as showing that up to two years living in the UK was not enough to impact on the NNSs' L2 MC (Azuma, 2005), and/or that time spent living in the UK had not impacted offline, written L2 MC, but may have brought gains for online, spoken L2 MC, undetectable by the MC Test Battery. A longitudinal study in this area, and research into the relationship between offline, written and online, spoken L2 MC was suggested.

Finally, a comparison of the correlations between MC-R and MC-P for 'low' (CEFR B1 or less), 'mid' (CEFR B2) and 'high' (CEFR C1 or above) NNS proficiency groups and the NS group showed a general downward correlational trend between modes (RQ6). The absence of significant correlations between MC-R and MC-P at higher L2 proficiency levels (i.e., CEFR C1 or above) and the NS level aligned with similar research in the area of L2 receptive and productive vocabulary knowledge (Henriksen, 2008), and the generally established finding that form-recognition is more closely associated with L2 proficiency than form-recall (Laufer & Goldstein, 2004; Schmitt, 2010). While it was speculated that these findings indicate that in similar NNS samples, individual, creative and stylistic freedom might be more pronounced at higher L2 (and native) proficiency, and that L2ers gain more control and become more selective (and varied) in the metaphors they produce at higher L2 proficiency levels, this requires empirical verification. Nonetheless, to the extent that L2 receptive and metaphoric competence behaves like L2 receptive and productive vocabulary knowledge, research would suggest that classroom tasks

centred on fostering the development of productive knowledge would bring greater gains for both L2 receptive and productive metaphoric competence than those designed to increase receptive knowledge (Webb, 2005).

11.3 Limitations and future research

It is possible to identify limitations in the present study concerning the generalisability of findings, the application of theory, and methodology. These are presented below alongside suggestions for future research to accompany those already mentioned during the discussions.

11.3.1 …related to the generalisability of findings

First and most significantly, since analyses were not replicated, it remains to be seen whether the L2 MC, vocabulary and proficiency tests would yield the same reliability coefficients, underlying factors, and predictive abilities if administered to a comparable (but different) sample of L1 Chinese NNSs of English. It would be particularly interesting to observe whether the same factor structure emerged, both for verifying the hypothesised (sub)competences of L2 MC, and also for determining the extent to which the 'stable' loading variables (as suggested by the bootstrap analysis) would re-emerge with external replication (T. J. B. Kline, 2005; Zientel & Thompson, 2007).

Second, although there were several methodological and investigative reasons to use NSs in the present study, the unequal match in terms of age and occupation might comprise a limitation on the validity of findings. Although both NNS and NS samples were argued to comprise speakers of higher intellectual profiles (Hulstijn, 2011, 2012) and the NSs were thought to be fairly representative of UK citizens of their socioeconomic status (SES) and thus able to provide generalisable data on so-called 'NS norms', their average age and age range (M = 39.7, SD = 16.3) was older and wider than that of the NNSs (M = 22.9, SD = 2.6), and unlike the NNSs, they were not all enrolled in or currently engaged in university study. A replication of the study on similar samples might therefore seek to bring the age ranges and occupations of the two groups more in line with one another.

A third limitation is that results pertain only to (university age) L1 Chinese learners of English. This caveat was also noted by Azuma (2005), who pointed out that her findings would not necessarily transfer to other L2 English speakers from very different cultural backgrounds. Suggestions about where findings may generalise to NNSs with other L1 backgrounds were made at a few points in the study (e.g., section 8.3.2.2 phrasal verbs), however, a new set of research questions is needed to investigate the extent to which factor structures and regression predictors would replicate on NNSs from either homogenous or mixed L1 backgrounds other than L1 Chinese. Such an investigation would complement existing research on culturally marked

metaphors (e.g., Liu & Xiao, 2013) and on potential areas of cross-linguistic influence.

Fourth, the present study dealt only with metaphor in the written mode, and in language. However, the extent to which the findings would extend to L2 MC in the spoken mode (e.g., as measured by Littlemore, 2001) or metaphor in visual modes is unknown.

Finally, researchers advocating an ELF perspective might also argue that notions of L1 and L2 MC have limited generalisability in an increasingly globalised and internet-connected world. In response to this, future investigation may need to place more emphasis on metaphoric competence in an English as a lingua franca context. For instance, studies might enquire whether MC in an ELF context is related to the ability to adapt to the metaphoric competence of an interlocutor or take into account an interlocutor's L1 (Pitzl, 2009, 2016). This is an important extension of the present study, which did not examine oral interaction in the 'wild'. Since metaphor use in ELF *does* tend to be measured in oral interaction (often via corpus methods), it is difficult at present to see how the MC Test Battery might fit into such research.

11.3.2 …related to the application of theory

First, the study is limited in its treatment of various figurative tropes and metaphor theories. While the role of idiom, simile (i.e., direct metaphor) and to a lesser extent, metonymy, in test items were discussed periodically, it remains to be seen whether (for instance) certain tropes were more difficult for NNSs, or made items better or worse at discriminating between high and low ability test takers. The fact that participants were informed they were signing up to a 'metaphoric competence' study probably means most of the metaphors they produced were deliberate. However, deliberate metaphor theory as such (Steen, 2008, 2011a, 2011b, 2013, 2015, 2016, 2017) did not form any part of the analyses. Given the growing debates around the empirical validity of deliberate metaphor theory and emerging identification procedures (Reijnierse et al., under review), future research might viably explore, whether (for instance) some factors identified involve more prototypically deliberate metaphors, or whether the deliberateness of metaphors affects how well the WAT, OOPT (overall) and other measures predict variance in L2 MC.

Second, although the grammatical accuracy of metaphor productions was not assessed, a further investigation could seek to identify patterns of syntactic errors or avoidance in the data, and what this reveals about metaphoric competence at the phraseological level (Philip, 2010). Such findings could be compared with the structural features of Chinese to detect possible areas of cross-linguistic influence (though a second L2 group would be required to properly investigate this).

Finally, another line of further research concerns the possible effects of various interventions. Although interventions involving aspects such as metaphorical phrasal verbs are

fraught with potential problems (section 11.4) the effect of different teaching techniques, approaches to noticing, giving feedback, and investigation into which subcomponents of metaphoric competence are sensitive to instruction may all prove pedagogically useful.

11.3.3 …related to methodology

Although it was argued that the use of elicitation methods allowed for better control (compared with naturalistic methods) over specific metaphors and functions of metaphor targeted, the use of elicitation methods might lead to criticism that the metaphoric competence engaged in lacks real-world validity. To this end, it is important to establish the extent to which Low (1988) and Littlemore and Low's (2006) metaphor-related skill and (sub)competences form part of natural, day-to-day communication. Future research might seek to identify instances of these in the 'wild' via ethnographic methods, or analysis of corpora such as VOICE (Seidlhofer et al., 2013).

Second, despite numerous steps to ensure that the EFA results were as robust as possible, EFA is by nature an exploratory tool. Once there is sufficient evidence that factor structures can be replicated in similar samples, Confirmatory Factor Analysis and Structural Equation Modelling could be used to investigate the conceptual independence of and causal relationships between factors, or suggest amendments to the theoretical model.

Third, analysis of change in correlation strength of MC-R and MC-P at different L2 proficiency levels might be criticised on methodological grounds. Although the practice of classifying NNSs into 'low', 'mid' and 'high' proficiency groups has been commonplace in research on the correlation strength change between L2 vocabulary size and depth as L2 proficiency increases (e.g., Henriksen, 2008; Noro, 2002; Nurweni & Read, 1999; Schmitt & Meara, 1997), some feel it is unnecessarily reductive. Citing J. Cohen (1983), Plonsky and Oswald (2017, p. 583) bemoan that "taking a continuous variable and artificially dividing it into two or more groups is a serious mistake, because you lose all the underlying continuous information for no good reason". Although in the present study it is argued that the L2 proficiency gap between 'low' (CEFR B1 or less) and 'high' (CEFR C1 or above) NNS groups means that the downward trend observed is likely to have been valid, for future enquiry it would be preferable to select a method which retains L2 proficiency as a continuous variable.

Fourth, although a systematic approach to scoring was taken, the fact that different raters did not always agree constitutes an inherent limitation. Future studies might address this problem by emulating Beaty and Silvia (2013) and entering different raters' scores for each MC test as different variables into the EFA. If two different raters' scores for one MC test loaded on the same factor, this could be taken as further validation of (sub)constructs.

A final methodological limitation is that findings only pertain to the different ways in which L2 vocabulary size and depth, and L2 proficiency were measured. It is unknown whether

the same findings would have been produced had vocabulary size been measured using the Eurocentres Vocabulary Test (Meara & Jones, 1990), Vocabulary Levels Test (Nation, 1983), or controlled productive vocabulary size Laufer and Nation (1999); or vocabulary depth using the Vocabulary Knowledge Scale (Wesche & Paribakht, 1996) or 1K-Vocabulary Depth Test (Richard, 2011). Future research might explore differences in the abilities of these tests to predict L2 receptive and productive metaphoric competence.

11.4 Implications for the EFL classroom

Because the present study had a research and testing focus, and since the limitations and further research agendas presented above have highlighted important generalisability, theory and methodology related issues, it is not possible to give anything beyond a tentative list of general teaching implications from the present findings.

The teaching implications from **Analysis 1: The development and reliability of the MC Test Battery, descriptive statistics** were twofold. First, Littlemore and Low's (2006a) general criticism of the use of multiple-choice questions in EFL textbooks and suggestion that multiple-choice activities require input from the teacher and the use of diagrams implies that such items in the MC Test Battery should be applied with care by EFL teachers. For instance, rather than having learners work through numerous multiple-choice questions individually, teachers could encourage students to debate the appropriateness of *best* answer and distractor options or work in groups to identify why certain options are more or less acceptable. To this end, the distractor analysis would help teachers anticipate which options are likely to elicit more disagreement and debate over acceptability between students.

A second implication concerned Test 8-Idiom Extension-P. Contrary to what Littlemore and Low (2006) found, as a group, the NNSs were generally capable of extending idioms. However, the fact that both NNSs and NSs struggled to think of extensions for several test items warns that teachers need to be aware that not all idioms are equally extendable.

The main teaching implication from **Analysis 2: Metaphoric and other (sub)competences uncovered**, was that **English Grammatical Metaphoric Competence** may be a particularly fruitful area to target from a pedagogical perspective, depending on the extent that learners want to have nativelike forms at their disposal. Concerning phrasal verbs, the effectiveness of highlighting underlying conceptual metaphors for language gains varies greatly according to how much these metaphors make sense to learners, and factors such as their age and cognitive style. Although drawing learners' attention to, for instance, the conceptual underpinnings of words and collocations has shown some learning advantage (Boers, 2000), the fact that single prepositions such as *up* can denote concepts as diverse as HAPPINESS, OBSTRUCTION, and COMPLETION (Kövecses & Szabó, 1996) may be of little help for learners

seeking hard and fast rules to aid the development of their L2 metaphoric competence. Whether or not the problems that the present study's NNSs had with phrasal verb particles could be alleviated by a conceptual metaphor-based intervention remains to be seen, but it seems unlikely that this approach will be suitable for all learners (MacArthur, 2010; Nacey, 2013).

For researchers and teachers seeking to exploit the metaphorical underpinnings of phrasal verbs, Littlemore and Low (2006a) suggested that if a meaning can be found, it should probably be taught. In this respect, the present study's findings could be useful for reference on which phrasal verbs (and particles) L1 Chinese learners found particularly challenging, and the types of distractors that lured them. However, in using these data for instructional purposes, EFL teachers have to be careful not to accidentally help learners acquire distractors (as was the case in Boers et al., 2014).

Another difficult teaching issue stemming from Analysis 2 is how to give learners feedback on the interpretations they arrive at and metaphors they produce. Several studies on corrective feedback (e.g., Goo & Mackey, 2013; Lyster, Saito, & Sato, 2013) suggest that in order for learners to attend to the differences between their production and a more acceptable one, it would be best for teachers to have learners supply the correction themselves. This aligns with Littlemore and Low (2006a), who emphasise both learner autonomy and the authority of the teacher, corpus or other learner resource to arbitrate between the metaphors that learners *do* and *should* produce. However, when it comes to metaphoric competence, teacher authority is problematic. As the NS data showed, in many cases, the NSs disagreed substantially on the extent to which a metaphor or particular expression of a Vehicle term is acceptable. In this view, students would only be getting from the teacher one of many subjective opinions. Moreover, imposition of any authority on metaphoric competence other than communicative usefulness is likely to be contentious for ELF advocates, who would argue for a more learner-driven or discourse context-driven approach to feedback on the 'appropriateness' of learners' metaphor productions. By implication, such an approach would necessitate more teacher tolerance to NNS varieties. If Test-3-Vehicle Acceptability-R and Test 4-Topic/Vehicle-R are to be used for such purposes, radically different scoring approaches would be needed.

To the extent that findings generalise to similar samples and learners with different L1s, the major teaching implication from **Analysis 3: Relationships between L2 metaphoric competence, vocabulary knowledge, general proficiency, age of starting to learn English and time spent living in the UK** is that both receptive and productive metaphoric competence (but particularly productive) will be most improved by tasks aimed at strengthening connections between representations within learners' lexical networks (i.e., knowledge measured by the WAT). Furthermore, to the extent that L2 metaphoric competence develops like L2 vocabulary knowledge, research suggests that classroom tasks fostering the development of productive

knowledge are likely to bring greater gains for both L2 receptive and productive metaphoric competence than those designed to increase receptive knowledge (Webb, 2005).

11.5 Contributions of the study

In conclusion, the present study has made several important contributions to the field of L2 metaphoric competence research.

First, despite Low (1988) and Littlemore and Low's (2006a, 2006b) metaphor-related skills and (sub)competences existing as theoretical constructs in the literature for 29 and 11 years respectively, they had never been elicited or used to develop tests. The present study changed that, not only by identifying challenges in the extent to which these metaphor-related skills and (sub)competences can be reliably elicited and measured, but also by providing substantive findings on factors underlying L2 metaphoric competence (as measured), and the ability of L2 vocabulary knowledge and general proficiency to predict L2 metaphoric competence. The practical outcome is both a wealth of substantive data and a comprehensively developed and refined MC Test Battery than can be used for further research, testing and teaching.

The second main contribution of the present study is the fact that it comprises a series of firsts. To the best of the present author's knowledge, the present study is the first known attempt to: 1) systematically review reliability in metaphoric competence research, thus showing that previously observed mixed instrument reliability (e.g., Littlemore, 2001) is a field-wide issue and that L2 metaphoric competence tends to yield lower reliability than the SLA field in general (Plonsky & Derrick, 2016); 2) apply structural Equation Modelling approaches seen in L1 metaphoric competence research (Beaty & Silvia, 2013; H. R. Pollio, 1977; H. R. Pollio & Smith, 1980; Silvia & Beaty, 2012) to analyse metaphoric competence in the L2, while also using bootstrapping techniques to gather empirical estimates of internal replicability (Davison & Hinkley, 1997; T. J. B. Kline, 2005; LaFlair et al., 2015; Plonsky et al., 2015; Zientel & Thompson, 2007); 3) determine how well vocabulary size and depth compare as predictors of L2 metaphoric competence (despite similar studies on their strength as predictors of L2 reading comprehension, e.g., Qian, 2002); 4) compare the relative power of L2 vocabulary knowledge and general proficiency, age of starting to learn English and time spent living in the UK as predictors in a combined model, thus corroborating relationships between these variables and receptive and productive metaphoric competence observed via correlation analyses (Aleshtar & Dowlatabadi, 2014; Azuma, 2005; Zhao et al., 2014); and 5) investigate the change in correlation strength between L2 receptive and productive metaphoric competence at different L2 proficiency levels, thus refining our understanding of these two modes as generally weakly correlated in 'intermediate' learners (Azuma, 2005; Littlemore, 2001).

The third main contribution of the present study is through its trial, testing and advocating of more robust approaches to the study of L2 metaphoric competence. Apart from a few exceptions (e.g., Littlemore, 2001), L2 metaphoric competence tests used in past research (e.g., Aleshtar & Dowlatabadi, 2014; Azuma, 2005; NourMohamadi, 2010; Zhao et al., 2014) have on the whole been highly limited in scope and their attention to data cleaning and test refinement has been minimal. Particularly in this respect, it is hoped that the methods used in the present study have allowed for us to get more of a grasp on this inherently 'slippery' construct.

Appendix A Rater training materials and scoring criteria for limited produced responses

Key terms

Metaphor - One thing (e.g., language, concept, image) treated as if it were, in some way, another

Linguistic metaphor - Language that can be considered metaphorical, for example "he's really got a screw loose!"

Vehicle (term) - A lexical item (or items) with an interpretation incongruous with the surrounding discourse thus signalling the presence of a linguistic metaphor, for example the words "got a screw loose!" in the linguistic metaphor above.

Topic – What a discussion is really about, for example "he's really got a screw loose!" is about *his craziness* (though the word 'crazy' might not be used explicitly)

Conceptual attributes of a Vehicle term or Topic - Any Vehicle or Topic contains conceptual attributes. These are characteristics or features associated with it. In the 'screw loose' example,

The Vehicle attribute of '[screws used for] securing physical structures' maps onto the Topic attribute of 'mental and behavioural unpredictability'

Conceptual metaphor - An underlying metaphorical concept that linguistic metaphors (are said to) point to, for example the linguistic metaphor above suggesting THE MIND IS A MACHINE

Source and target domains - The conceptual domains that a conceptual metaphor engages, for example MACHINES (source domain) used to understand THE MIND (target domain)

Recognised saying (formulaic sequence)

For the purposes of this research, a recognised saying (formulaic sequence) is a combination of words that you might often see together in a relatively fixed combination, for example, 'where there's a will there's a way'

Critical item - The part (e.g., word) of the test item sentence that the participant is being tested on.

Critical part of a response - The part of the response that is important to assessing whether the respondent has understood something or produced an appropriate response.

Practice examples

Please use the scoring criteria to score the following responses:

Test	Question	Response	Meaning score[98]
2	1a) The news lifted her spirits… **This means**	*The news made her happy.*	(1)
4	Q9. New products at the end of a long production process are the _____of large companies. **Please complete the analogy:**	*goods*	(0)
5	Q10. **Speaker A:** I get the feeling that this project is becoming complicated **Speaker B:** why's that? **Speaker A:** Well, at the meetings, everybody wants to take the lead and push their ideas. I just feel that because there's a lot of people involved, that it's having a negative effect on progress. **Speaker B:** Well, you know what they say, _____. **Please write an appropriate phrase/expression to finish the dialogue:**	*A great man cannot brook a rival.*	(1)
6	Q9. The stomach functions like _____. **Please type something suitable to describe to a child what the stomach functions like:**	*factory*	(1)
7	Q12. Your colleague Michelle is very unkind and nasty. She spreads untrue rumours about people in the office. **Please complete the comment to show how you feel about Michelle:** Michelle is about a nice as_____.	*stepmother of snow white*	(2)
8	Q12. (Original idiom: it's raining cats and dogs = it's raining a lot) Extended idiom: It's been raining cats and dogs for so long that _____. **Please extend the idiom:**	*it was full of animals outside*	(1)
9	Q10 **Peter:** Have you heard, the wizard has done his magic again? I mean the secret magic award **You:** oh yes, I heard Mr magic is due to be _____. **Please write responses in 'code' that keep the conversation with Peter going:**	*be anointed as a grand high wizard*	(2)

[98] Numbers in parenthesis denote the researcher's intended *correct* score out of 2.

Scoring criteria: MC Test Battery limited production responses

Test 2-Metaphor Layering-R (scored 1 or 0)

Question	Score	Criteria for scoring meaning	Example response[99]
Q1. a) The news **lifted** her spirits This means…	1	The respondent has explained the meaning as becoming happier, cheering up, encouragement, being heartened, an uplifting feeling, alleviating stress, improving mood or has used another description synonymous with any of these.	*the news cheered her up*
	0	The respondent has: (a) explained that this means that the news has motivated or excited her, made her feel happy (as opposed to happier), energetic, inspired or has used another description synonymous with any of these; (b) provided another incorrect meaning; (c) used words from the critical item in their explanation in a way that does not allow for assessment of their understanding of meaning; (d) not attempted the question	*a) the news made her happy* *b) the new is encouraging her to do sth* *c) her spirits was lifted by the news*
Q2. a) She treated us in a **cold** way This means…	1	The respondent has explained that this means that she was unfriendly, unsympathetic, uncaring, lacking feeling, distant, aloof, unpleasant, not warm, indifferent, unwelcoming, abrupt, rude, not kind, not well, that she behaved in a cool manner or has used another word synonymous with any of these.	*she treated us indifferently*
	0	The respondent has: (a) explained that this means bad, unhelpful or hostile treatment, a lack of enthusiasm, or has used another word synonymous with bad or unhelpful to describe the treatment; (b) provided an incorrect meaning; (c) used words from the critical item in their explanation in a way that does not allow for assessment of their understanding of meaning, or (d) not attempted the question	*a) she treated us very badly* *b) she is not an easygoing person* *c) her treatment of us was cold*
Q3. a) They will want to get married sometime in the **distant** future This means…	1	The respondent has correctly mentioned what the implications of the statement are for the future (i.e., that the couple will get married, but not for a while) and not just described the present state of affairs. The answer might include the wording someday, sooner or later, sometime, to come, one day, far off or similar.	*they will not get married very soon, but in the future which is far away from now.*
	0	The respondent has: (a) explained that this means that the couple do not want to marry now/will want to marry/ there is little chance of marriage but with no mention of the future; (b) made an unwarranted inference; (c) provided an incorrect meaning, (d) used words from the critical item in their explanation in a way that does not allow for assessment of their understanding of meaning, or e) not attempted the question	*a) they don't want to get married now* *b) they are now single, but in the future, they will get married.* *c) they won't want me get married* *d) They desire marriage in the distant future time*
Q4. a) He has a **fiery** temper This means…	1	The respondent has explained that this means that he is bad tempered, not good-tempered, irritable, moody, touchy, short tempered, sensitive, temperamental, fractious, hot tempered, easily annoyed, enraged, irascible, has extreme anger or has used another word synonymous with any of these.	*he gets angry easily*
	0	The respondent has: (a) explained that this means being not nice, rude, impatient, has provided a general comment about the person being bad, introduced irrelevant meanings or used another word synonymous with any of these; (b) provided an incorrect meaning; (c) used words from the critical item in their explanation in a way that does not allow for assessment of their understanding of meaning, or (d) not attempted the question	*a) his temper is not good enough* *b) he gets excited easily* *c) his temper is kind of fiery*
Q5. a) The conscience is **man's compass** This means…	1	The respondent has explained the conceptual link between a compass and the conscience in terms of the latter being a guide towards the correct (moral) direction	*The conscience tells a man what is the right thing to do.*
	0	The respondent has explained that the conscience is: (a) involved in thought and decision making but with no mention of it being a guide or help; (b) linked to actions but with no mention of right or wrong; (c) valuable or important to man (people) but with no mention of why, or has (d) provided an incorrect meaning, e) used words from the critical item in their explanation in a way that does not allow for assessment of their understanding of meaning, or f) not attempted the question	*a) thinking before doing* *b) man need conscience to do rational things* *c) the conscience is very important to man* *d) the conscience is within man's control* *e) the conscience becomes man's compass*
	1	The respondent has explained that this means that TV provides no real intellectual sustenance or is a mindless, repetitive, monotonous	*TV passes time but provides no sustenance*

[99] Unless otherwise indicated, all examples are real NNS or NS productions from the pilot or main studies.

Q. 6 a) TV is **chewing gum** for the eyes This means…		activity, form of negative stimulation, or used for killing or wasting time.	
	0	The respondent has: (a) explained that this means that TV is bad, harmful, boring, entertaining, time consuming, positive stimulation or addictive; (b) provided an incorrect meaning; (c) used words from the critical item in their explanation in a way that does not allow for assessment of their understanding of meaning, or (d) not attempted the question	*a) TV is bad for the eyes* *b) people needs to watch TV* *c) TV is like chewing gum for a person's eyes*

Test 4-Topic/Vehicle-P (scored 2, 1 or 0)

Score	Meaning – Criteria for scoring productive responses
2	The response results in the formation of an analogy that is clearly understandable and makes logical sense. To award '2', the scorer must be able to identify: a) Which conceptual attributes of the response (i.e., Vehicle term) are relevant b) Which generalisable conceptual attributes of the entity being described (i.e., Topic) are relevant c) How these relate to one another to form the analogy
1	The response results in the formation of an analogy that is somewhat understandable and makes some logical sense, but there are problems with regard to conducting the processes described in a), b) and/or c) above.
0	Either the response results in an analogy that is not understandable, does not make logical sense, the result is not an analogy but rather a literal description, or no response is given.

Test item	Score	Example response	Example scorer justification
Q7. The sales team are the _____ of the organisation	2	*hunters*	Hunters bring home a kill to sustain a society. Similarly, a sales team metaphorically 'brings home' income to sustain an organisation.
	1	*skin*	Skin is what a person looking at the body mostly sees (along with hair, eyes, etc.). Similarly, the sales team are what the customer/buyer mostly sees of an organisation. There are significant attributes of skin, however, that don't easily metaphorically relate to sales teams: skin appears fixed on the body but is also elastic, it is relatively thin, it can tan, blush, change colour, be cut and so on.
	0	*a) window* *b) main part*	a) It is difficult to perceive which features of window relate to which features of sales teams. b) This forms a literal description rather than an analogy
Q8. Alcohol is the _____ of the drunk person	2	*fuel*	Fuel gives energy and 'life' to an engine. Similarly, alcohol gives energy and 'life' to a drinker.
	1	*bread*	Bread is a basic foodstuff consumed daily by many people. Similarly, alcohol might be regarded as basic and consumed daily if the drunk person is an alcoholic. There are significant attributes of bread, however, that don't easily map onto attributes of alcohol: bread is solid and crumbly, it is not usually a main dish, though a staple is not regarded as a high energy, enticing or metaphorically 'intoxicating' foodstuff.
	0	*a) sore* *b) reason*	a) it is difficult to perceive which features of sore relate to which features of alcohol b) This makes for a literal description rather than an analogy
Q9.The outside walls are the _____ of the building	2	*skin*	Skin forms the outer layer of the body. Similarly, the walls of a building form its outer layer.
	1	*clothes*	Clothes are worn on top of the body and form an outer layer (outside the body). Similarly, the walls form the outer layer of a building. There are significant attributes of clothes, however, that don't easily map onto attributes of outside walls: clothes are often creased, they appear soft, move with the body, contain embellishments, and so on.
	0	*a) arms* *b) protection*	a) it is difficult to perceive which features of arms relate to which features of outside walls b) This makes for a literal description rather than an analogy
Q10. Killer whales are the _____ of the sea	2	*wolves*	Wolves hunt in packs, similarly killer whales hunt as a team.
	1	*Obama*	Obama was the US President when the test was administered (2015). Similarly, killer whales are top of their food chain. There are significant attributes of Obama, however, that don't easily map onto attributes of killer whales: Obama is perhaps primarily known as being the first black US president, he is not known for aggressive international policy (compared to his predecessors) and won the 2009 Nobel peace prize.
	0	*a) enemy* *b) animals*	a) This answer makes it seem as if killer whales are at enmity with the sea, which doesn't make sense. b) This makes for a literal description rather than an analogy
Q11. Volcanoes are the _____ of the earth	2	*pimples*	Pimples are cone shaped bumps which release pus. Similarly, volcanoes are cone shaped formations that release magma.
	1	*scar*	Scars appear as a ridge or bump on the skin. Similarly, volcanoes form a bump on the earth's surface. There are significant attributes of scars,

			however, that don't easily map onto attributes of volcanoes: scars are typically sealed and don't emit liquid, scars are typically long rather than circular bumps and so on.
	0	*a) lymph* *b) normal status*	a) it is difficult to perceive which features of lymph relate to which features of volcanoes b) This makes for a literal description rather than an analogy
Q12. Chemical elements are the _____ of life	2	*building blocks*	Building blocks are pieces that fit together to form bigger more complex structures, similarly chemical elements combine to form more complex structures.
	1	*flavour*	Chemical elements are basic countable structures that bond to make more complex structures (including living organisms). Similarly, flavour can refer to flavouring, an ingredient that mixes with other ingredients in food. There are significant attributes of flavour, however, that don't easily map onto attributes of chemical elements: a flavour is a perception of taste and not something physical, a flavouring is often not a central component of a meal.
	0	*a) sweet poison* *b) basic components*	a) it is difficult to perceive which features of sweet poison relate to which features of chemical elements b) This makes for a literal description rather than an analogy
Q7. The main argument is the _____ of the essay	2	*meat*	Meat is regarded as a main component of a meal. Similarly, the main argument forms the main component of an essay.
	1	*key*	The key is an essentially item for opening a door. Similarly, the main argument is a vital part of a written composition.
	0	*a) eye* *b) topic*	a) it is difficult to perceive which features of an eye map onto which features of a main argument b) This makes for a literal description rather than an analogy
Q8. The CCTV cameras are the _____of the building.	2	*eyes*	Eyes perceive everything in front of them, similarly CCTC cameras record everything in front of them.
	1	*housekeeper*	A housekeeper takes care of a building. CCTV cameras help take care of a building (and its contents). There are significant attributes of housekeepers, however, that don't easily map onto attributes of CCTV cameras: housekeepers actively clean, CCTV cameras simply help detect problems and so on.
	0	*a) pillar* *b) security assistance*	a) it is difficult to perceive which features of a pillar map onto which features of CCTV cameras b) This makes for a literal description rather than an analogy
Q9. New products at the end of a long production process are the _____of large companies	2	*fruit*	Fruit is the result of natural processes and (often) human input over a certain period, this is similar to companies producing new products over a certain period.
	1	*dessert*	Desserts are generally desirable and come at the end of a meal. Similarly, new products come at the end of a long production process and are desired by customers. There are significant attributes of desserts, however, that don't easily map onto attributes of new products: desserts are not the main sustenance of a meal and are often regarded as treats rather than necessary for staving hunger, desserts can be unhealthy for a person and so on.
	0	*a) warrior* *b) goods*	a) it is difficult to perceive which features of a warrior map onto which features of new products b) This makes for a literal description rather than an analogy
Q10. This park is the _____of our city	2	*lungs*	Lungs convert oxygen to carbon dioxide in order to sustain a body, this is similar to a park, which helps sustain fresh air by converting carbon dioxide back into oxygen.
	1	*liver*	The liver detoxifies metabolites. Similarly, a park cleanses a city's air. However, the analogy is somewhat odd because the liver deals with fluids (rather than gas).
	0	*a) bed* *b) landmark*	a) it is difficult to perceive which features of a bed map onto which features of a park. b) This makes for a literal description rather than an analogy
Q11. The bee hive is the _____ of the animal kingdom	2	*airport*	the airport is a hub for air traffic, similarly, the bee hive is a base for bees (which fly).
	1	*bomb*	A bomb is an intricate device containing wires and other components. Similarly, a bee hive contains intricate tunnels and chambers. There are significant attributes of bombs, however, that don't easily map onto attributes of bee hives: bombs explode, bombs have a regular tick and can often be detonated remotely and so on.
	0	*a) lip* *b) epitome*	a) it is difficult to perceive which features of a lip map onto which features of a bee hive b) This makes for a literal description rather than an analogy
Q12. The company's internal mail team are the _____ of the organisation	2	*blood*	Blood flows around the body supplying organs with oxygen, similarly, the company's internal mail team move around the offices and floors of an office block, supplying employees with mail.
	1	*conveyor belt*	Conveyor belts carry items along a production line. Similarly, an internal mail team deliver mail from workstation to workstation. There are

			significant attributes of conveyor belts, however, that don't easily map onto attributes of internal mail teams. Conveyor belts are long belts that move in one direction, they carry items that get increasingly put together and so on.
	0	*a) nerves* *b) transportation*	a) it is difficult to perceive which features of nerves map onto which features of an internal mail team b) This makes for a literal description rather than an analogy

Test 5-Topic Transition-P (scored 2, 1 or 0)

Score	Meaning – criteria for scoring productive responses
2	The response finishes the dialogue appropriately by way of some proverbial advice or a proverbial summary of the other speaker's situation and is either: a) A recognisable attempt at a metaphorical or literal phrase in Macmillan dictionary (type A) b) Clearly idiomatic (i.e., has an overall meaning different from the sum of the individual words) or a phrase found in popular culture (e.g., song title, slogan) but is not in Macmillan dictionary (type B) c) An expression that contains a metaphor in a critical part (type C)
1	The response finishes the dialogue somewhat appropriately by way of some proverbial advice or a proverbial summary of the other speaker's situation and is either: a) A recognisable attempt at a metaphorical or literal phrase in Macmillan dictionary (type A) b) Clearly idiomatic (i.e., has an overall meaning different from the sum of the individual words) or a phrase found in popular culture (e.g., song title, slogan) but is not in Macmillan dictionary (type B) c) An expression that contains a metaphor in a critical part (type C)
0	Either the response is type A, B or C but does not make logical sense in context (type D), or is literal advice or a literal summary that is not a Macmillan phrase (type E), or no response is given (type F)

Test item	Score	Example response	Example comments
Q7. **Speaker A:** You know, it's funny when I think about my dad. **Speaker B:** Why's that? **Speaker A:** We have exactly the same habits. We both like to get up early, enjoy watching history documentaries, and I suppose we're both kind of quiet and passive most of the time. **Speaker B:** well, you know what they say, _____ **Please write an appropriate phrase/expression to finish the dialogue:**	2	*a) Like father, like son* *b) the apple never falls far from the tree.* *c) children are just carbon copies of their parents*	a) The response is proverbial and is a (literal) phrase in Macmillan (type A) b) The response is proverbial and clearly idiomatic but is not in Macmillan (Type B) b) The response is appropriate, a critical part of it contains a metaphor (i.e., carbon copies) (type C)
	1	*birds of a feather flock together*	The response is somewhat appropriate since the father and child have the same tastes and are found together, but it is not ideal since it does not link the fact that the child came from the father (type A)
	0	*a) blood is thicker than water* *b) sons are alike dads*	a) The assertion that family ties are more important than ties made among friends is irrelevant here (type D) b) This is a literal summary of the facts and is neither a Macmillan phrase nor a metaphor (type E)
Q8. **Speaker A:** I've lived all over the world. I was born in India but grew up in Germany. I spent some time in the USA and Australia and have been in the UK for just six months. **Speaker B:** So where do you consider to be home? **Speaker A:** Difficult question! But I suppose, when I think of my wife, I don't mind where I live as long as it's with her. **Speaker B:** That's wonderful, you know what they say, _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*a) home is where the heart is* *b) home is where love dwells*	a) The response is proverbial and clearly idiomatic but is not in Macmillan (type B) b) The response is appropriate, a critical part of it contains a metaphor (i.e., love dwells) (type C)
	1	*there is no place like home*	The response is somewhat appropriate since the speaker clearly values 'home', but it is not ideal since it does not link the fact that 'home' finds true definition in relation to the speaker's wife (type B)
	0	*a) love me, love my dog*	a) The assertion that one must love the speaker's dog if they love the speaker is irrelevant here (type D)
Q9. **Speaker A:** I'm so glad we double checked the proposal for the product design before sending it to the manufacturers	2	*a) a stitch in time saves nine* *b) a small leak will sink a great ship*	a) The response is a common proverbial saying and is in Macmillan (type A)

Speaker B: Why, was there something wrong in the plan? **Speaker A:** Very much so! In one section we had specified completely the wrong component! If that had gone unnoticed, in three months we would be spending tens of thousands on fixing the problem. **Speaker B:** Good that you spotted it, you know what they say, _____. **Please write an appropriate phrase/expression to finish the dialogue:**			b) The response is proverbial and clearly idiomatic but is not in Macmillan (type B)
	1	*wisdom comes by suffering*	The response is somewhat appropriate since the speaker would clearly learn from the mistake if it had been left unfixed and had caused problems, but it is not ideal since it does not focus on the positive aspects of taking preventative action (type B or C)
	0	*a) Being casino always can help to reduce the risk* *b) carelessness might lead to a huge problem*	a) It is unclear what 'being casino' means (type D) b) This is a literal summary of the facts and not a recognised saying, or not metaphorical (type E)
Q10. **Speaker A:** I get the feeling that this project is becoming complicated **Speaker B:** why's that? **Speaker A:** Well, at the meetings, everybody wants to take the lead and push their ideas. I just feel that because there's a lot of people involved, that it's having a negative effect on progress. **Speaker B:** Well, you know what they say, _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*a) too many cooks/chefs spoil the broth (soup)* *b) one nation can't have too queens*	a) The response is a common proverbial saying and is in Macmillan (type A) b) The response is proverbial and clearly idiomatic but is not in Macmillan (type B)
	1	*Thousands people, thousands brains.*	The response is a possible fit since one thousand brains would complicate things, but it is not ideal since one thousand brains could also be interpreted as a good thing in the sense that two heads are better than one (type B)
	0	*a) one rotten apple spoils the whole lot* *b) it's better to have one leader*	a) The idea of one small pollutant can ruin everything is not the point here (type B) b) This is a literal summary of the facts and not a recognised saying, or not metaphorical (type E)
Q11. **Speaker A:** I think I've lost all faith in mankind! **Speaker B:** that sounds a bit extreme, what happened? **Speaker A:** I just can't rely on anyone or anything. My friend keeps cancelling our meeting, my assistant at work didn't do what he's supposed to, the weather forecast said sun, it's raining! You name it, you can't predict anything! **Speaker B:** Well you know what they say about this life, _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*a) it never rains but it pours* *b) there are three things for sure: taxes, death and trouble*	a) The response is a common proverbial saying and is in Macmillan (type A) b) The response is proverbial and clearly idiomatic but is not in Macmillan (type B)
	1	*The good always comes in the end*	The response is a possible fit as optimistic encouragement but the conversation has a more pessimistic tone (type A)
	0	*a) we are just little part in the universe* *b) life is unpredictable*	a) The vastness of the universe and earth's relative insignificance are not relevant neither summarise the situation nor offer useful proverbial advice (type D) b) This is a literal summary of the facts and not a recognised saying, or not metaphorical (type E)
Q12. **Speaker A:** It's a shame that my brother and our friend Peter are not getting along well. **Speaker B:** What's the problem? **Speaker A:** Well, there's always been this tension between them, I just don't think they like each other very much. It's difficult because everyone has started to take sides. I like Peter very much, but if comes down to it, I have to support my brother, he's family. **Speaker B:** I understand, well you know what they say, _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*a) blood is thicker than water* *b) blood tie is stronger than friendship*	a) The response is a common proverbial saying. Blood stands for family ties, water stands for friendship ties (type A) b) The response is appropriate, a critical part of it contains a metaphor (i.e., blood tie) (type C)
	1	*everyone has two sides/everything [everything has two sides]*	The response is somewhat appropriate since this is a good example of a situation that has two sides, but it is not ideal since it does not comment on the fact that the speaker has chosen one of the sides according to a certain principle (type C)
	0	*a) gay is oke* *b) you have to do what you think is right and fair*	a) A statement in support of gay rights is irrelevant here (type D) b) This is a literal summary of the facts and not a recognised saying, or not metaphorical (type E)

Q7. **Speaker A:** We went to a small village in France last month **Speaker B:** oh that's great, what did you do? **Speaker A:** well, we really tried to enjoy the French culture and fit in with the locals. We drank fresh coffee and read a newspaper in the mornings, then ate lunch with wine and in the evening walked the streets listening to the live music. We really started to feel French! **Speaker B:** Great, well you know what they say, _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*when in Rome, do as the Romans do*	The response is a common proverbial saying (type A)
	1	*you only live once*	The response is somewhat appropriate since the speaker and his/her travelling companion(s) seem to have indulged in luxuries, but it is not ideal since it does not comment on the fact that they felt as though they were blending in with the locals (type B)
	0	*a) love me love my dog* *b) try to be a local*	a) The assertion that one must love the speaker's dog if they love the speaker is irrelevant here (type B) b) This is literal advice and not a recognised saying or a metaphorical utterance (type E)
Q8. **Speaker A:** Did I tell you that Sarah and I broke up last week? **Speaker B:** No! Oh that's so sad, how come? **Speaker A:** We just weren't right for each other. I'm so down; I just don't feel like I'll ever meet the right person **Speaker B:** I'm sure you will, I know Sarah was great but don't worry, you know what they say, _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*a) plenty more fish in the sea* *b) the world is so big, flowers are in everywhere*	a) The response is a common proverbial saying (type A) b) The response is proverbial and clearly idiomatic but is not in Macmillan (type B)
	1	*Beautiful recognizing and beautiful leaving.*	The response is somewhat appropriate if it is to mean that beauty can be found in all parts of the process of acquiring and losing a romantic partner, but it is not ideal because it does specifically give the other speaker hope of finding someone new (type C)
	0	*a) you will meet the right person at the right time*	a) This is literal advice and not a recognised saying or a metaphorical utterance (type E)
Q10. **Speaker A:** It was so difficult to find funding for my studies. I applied to seven different funding councils, all of them rejected me. I then looked at loan options and part time work. It was tough but I was so determined that I would find funding and start my studies. **Speaker B:** So did you have any success? **Speaker A:** Yes, I managed to get funding from the company I currently work for, they have a scheme for employees looking to continue their education. **Speaker B:** That's great, you know what they say, _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*a) where there's a will there's a way* *b) a man with a determined heart will beat every difficulty*	a) The response is a common proverbial saying (type A) b) The response is proverbial and clearly idiomatic but is not in Macmillan (type B)
	1	*After a storm comes a calm.*	The response is somewhat appropriate since the storm could represent difficulties and the calm the resolution, but it is not ideal because it doesn't comment on the speaker's determination and perseverance (type B)
	0	*a) Hard work makes perfect.* *b) this man will get some financial help from his current company*	a) the assertion that hard work makes perfect does not apply here, the situation is not about achieving perfection (type D) b) This is a literal summary and not a recognised saying or a metaphorical utterance (type E)
Speaker A: I'm so embarrassed **Speaker B:** why, what happened? **Speaker A:** I accidentally broke my colleague Peter's coffee mug at our office **Speaker B:** did he get angry? **Speaker A:** well, he was away on business yesterday, but he's back in later today, I'm so worried **Speaker B:** come on, don't worry, you know what they say _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*no use crying over spilt milk*	The response is proverbial and clearly idiomatic but is not in Macmillan (type B)
	1	*You will cross the bridge when you get to it.*	The response is somewhat appropriate since it suggests that the speaker can adjust his/her approach according to the developing situation, but it is not ideal because there are not problems that need to be dealt with in the meantime and the confrontation is inevitable rather than just possible (type A)
	0	*a) every road come to the Rome,* *b) just say it directly and it would be so easy.*	a) The assertion that there are many different routes to the same goal is not relevant advice here (type D) b) This is literal advice and not a recognised saying or a metaphorical utterance (type E)

Q11. **Speaker A:** I had an interesting dilemma the other day, my boss asked me to prepare a report over the weekend, to be ready for Monday morning. Unfortunately, I had a lot of stress with the furniture removal people on Saturday and I just forgot to do the report. It's not a good excuse but it's the truth. **Speaker B:** So what did you tell your boss? **Speaker A:** In the end I decided not to lie and that it's better to tell the truth and apologise. **Speaker B:** I agree, you know what they say _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*a) honesty's the best policy* *b) honesty is gold*	a) The response is proverbial and clearly idiomatic but is not in Macmillan (type B) b) The response is appropriate, a critical part of it contains a metaphor (i.e., honesty is gold) (type C)
	1	*A problem shared is a problem halved*	The response is somewhat appropriate since it suggests that by sharing the truth, the speaker can reduce the problem, but it is not ideal since its focus is not on coming clean, but rather on trying to solve a problem by sharing it with people who could contribute useful suggestions (type B)
	0	*a) consensus is man's campuss* *b) Tell truth is always the best*	a) The words taken as they are suggest a nonsensical proverb (the intention may have been 'the conscience is man's compass' but the production is markedly different) (type D) b) This is a literal summary and not a recognised saying or a metaphorical utterance (type E)
Q12. **Speaker A:** we had such a panic last week when the clients from Germany visited. My car broke down, the hotel had double booked, and we had a lot of employees away sick! **Speaker B:** that's terrible, so what happened? **Speaker A:** Thankfully, I was able to get a taxi and sort out everything with the hotel. It was actually OK with just a few staff in the office; it meant we weren't disturbed during our meeting. **Speaker B:** sounds crazy, but you know what they say _____. **Please write an appropriate phrase/expression to finish the dialogue:**	2	*a) all's well that ends well* *b) Rainbow will come out after a heavy rain.*	a) The response is a common proverbial saying (type A) b) The response is proverbial and clearly idiomatic but is not in Macmillan (type B)
	1	*life is a box of chocolate. You never know what is gonna happen next*	The response is somewhat appropriate since it points out the unpredictability of life, but it doesn't connect with the fact that everything turned out OK in the end (type B)
	0	*a) road will be straight when you come* *b) You can find a way to solve the problem finally.*	a) It is unclear what this proverb might be asserting (type D) b) This is a literal summary and not a recognised saying or a metaphorical utterance (type E)

Test 6-Heuristic-P (scored 2, 1 or 0)

Score	Meaning – criteria for scoring productive responses
2	The response forms a simile that suitably describes to a child what the item in question is, looks, sounds (etc.) like.
1	The response forms a simile that somewhat suitably describes to a child what the item in question is, looks, sounds (etc.) like but there may be some problems with the logic of the comparison.
0	Either the response results in a simile that is not understandable or does not make logical sense (see example (a)), or the result is not a simile but a literal comparison (see example (b)) or no response is given.

Test item	Score	Example response	Example scorer justification
Q7. Thunder sounds like_____. Please type in something suitable to describe to a child what thunder **sounds like:**	2	*a hundred horses running*	A hundred horses running and thunder sound similar in the sense that both emit a loud, low, rumbling sound. In addition, a child could understand the concept of a hundred horses running
	1	*my head hit on the wall*	This would produce a thud not dissimilar to thunder but the pain and specificity (i.e., my head) might mislead a child into thinking that thunder hurts them
	0	*a) knocking a gang* *b) long, low rumble (brontide)*[100]	a) Too much inference is required to work out what this could mean b) Too literal
Q8. Clouds function like_____. Please type something suitable to describe to a child what clouds **function like:**	2	*bags of water droplets*	Bags (of water droplets) and clouds function similarly in the sense that both hold things and gradually become heavier. A child could understand the concept of a bag (e.g., plastic bag) containing water

[100] All b) examples for 0 scores have been invented for the purpose of training the scorers.

	1	*the box where contain all toys*	There are similarities in terms of the box functioning as a storage unit but its angular shape makes this simile problematic
	0	*a) Boat on the river* *b) a visible mass of condensed watery vapour floating in the atmosphere*	a) It is difficult to see how a boat on a river functions like a cloud b) Too literal
Q9. The stomach functions like _____. Please type something suitable to describe to a child what the stomach **functions like:**	2	*a car fuel tank*	A fuel tank and a stomach function similarly in the sense that they both store a type of (liquid/liquidised) fuel and need topping up. A child would be able to understand the concept of a car fuel tank
	1	*fridge of your body*	There are similarities in terms of storage but the items in a fridge are not all liquid and do not contribute towards its working
	0	*a) fertilizer* *b) a muscular organ*	a) it is difficult to see how fertilizer functions like a stomach b) Too literal
Q10. The ozone layer functions like _____. Please type something suitable to describe to a child what the atmosphere/ozone layer **functions like:**	2	*protective bubble wrapping*	Protective bubble wrapping functions like the ozone layer in the sense that both form a protective layer and allow light through. A child could understand the concept of bubble wrapping
	1	*a protection umbralla*	There are similarities in terms of protection and, in part, curved shape, but problems in terms of the protrusion of the umbrella, the fact it only partly covers something (i.e., not from all angles)
	0	*a) vacuum machine* *b) a region of the earth's stratosphere that absorbs sunlight*	a) it is difficult to see how a vacuum machine functions like the atmosphere/ozone layer b) Too literal
Q11. The heart functions like_____. Please type something suitable to describe to a child what the heart **functions like:**	2	*a pump*	A pump and the heart function similarly in the sense that both contain 'pipes' and push air, liquid and so on into or around a system usually in a regular pumping manner. A child could understand the concept of a pump
	1	*battery*	There are similarities in terms of power supply but a battery does not project air/liquid and does not operate in a pumping manner
	0	*a) a straw* *b) a muscular organ*	it is difficult to see how the heart functions like a straw. This does not at all contain ideas about the heart's chambers and shape b) too literal
Q12. The roots of a plant function like_____. Please type something suitable to describe to a child what the roots of a plant **function like:**	2	*a ship's anchors*	A ship's anchors and the roots of a plant function similarly in the sense that both anchor their vessel to the ground. A child could understand the concept of a ship's anchors
	1	*feet*	There are similarities in terms of the relative location of feet and roots to their respective owners, but problems in the fact that feet allow mobility (i.e., walking) and do not anchor a person to the same spot.
	0	*a) a compass* *b) the organ of a plant lying beneath the soil*	a) it is difficult to see how a compass functions like the roots of a plant b) too literal
Q7. A disease in the body behaves like_____. Please type in something suitable to describe to a child what a disease in the body **behaves like:**	2	*an army on the attack*	An army on the attack and a disease behave similarly in the sense that both are comprised of many smaller entities that destroy things in their path but that can be fought against and defeated. A child could understand the concept of an army on the attack
	1	*bad things in refrigerator*	There are similarities in terms of both getting worse and worse but problems in terms of the fact that food going bad in a refrigerator does not affect the refrigerator's overall running and that the fridge does not have a system of making food go good again
	0	*a) teeth's worm* *b) an abnormal condition or disorder*	a) it is difficult to see how teeth's worm behave like a disease in the body
Q8. The brain works like_____. Please type	2	*a computer*	A computer and a brain work similarly in the sense that both contain many components, make powerful calculations and store information. A

in something suitable to describe to a child what the brain **works like:**			child could understand the concept of a computer.
	1	*steering wheel*	There are similarities in terms of each being a key source of control but problems in terms of the steering wheel being one of many parts that contribute to the operation of a car and something that does not operate automatically
	0	*a) the star.* *b) the centre of the nervous system*	a) it is difficult to see how the star works like the brain b) too literal
Q9. An electric current running through a wire is like_____. Please type in something suitable to describe to a child what an electric current running through a wire **is like:**	2	*water in a pipe*	Water in a pipe and electricity both flow, demonstrate a current and operate within sealed 'tubes'. A child could understand the concept of water in a pipe
	1	*cars on road*	There are similarities in terms of flow, but problems because a road is not a sealed 'tube'.
	0	*a) There is a storm in the river* *b) a flow of charge carried by moving electrons*	a) It is difficult to see how a storm in the river is like electricity in a pipe b) Too literal
Q10. Lava running down the side of a volcano moves like_____. Please type in something suitable to describe to a child what lava running down the side of a volcano **moves like:**	2	*syrup*	Syrup and lava both move in the save manner, are a similar colour roughly speaking. A child could understand the concept of syrup
	1	*the cheese from oven.*	There are similarities in terms of colour and melting effect but problems in terms of cheese's stringiness and the fact that it is much less runny that lava and so moves in a characteristically different way
	0	*a) a fast cheetah* *b) molten rock*	a) it is difficult to see how a fast cheetah moves like lava b) Too literal
Q11. Using letters to spell words is like _____. Please type in something suitable to describe to a child what using letters to spell words **is like:**	2	*fitting the pieces of a jigsaw puzzle together*	Fitting the pieces of a jigsaw puzzle together and using letters to spell words are similar in that both constitute small pieces that need to be arranged in a certain very specific order to achieve meaning. A child could understand the concept of fitting the pieces of a jigsaw puzzle together
	1	*making a meal with ingredients*	There are similarities in terms of parts contributing to a whole, but problems in terms of timescale and that ingredients often become unrecognisable once added to a meal.
	0	*a) use chopsticks to have meals* *b) comprising units of semiotic value to form a more complex entity*	a) it is difficult to see how using chopsticks to have meals is like using letters to spell words b) Too literal
Q12. Eye lids function like _____. Please type in something suitable to describe to a child what eye lids **function like:**	2	*shutters/blinds*	Shutters/blinds and eye lids function similarly in the sense that both shut out light from their interior
	1	*switch of the light*	There are similarities in terms of the result being the same (i.e., light on or off), but problems in terms of shape and location in relation to light source
	0	*a) the window of soul* *b) upper and lower folds of skin*	a) eye lids cannot be said to function like a window (which itself is transparent) b) Too literal

Test 7-Feelings-P (scored 2, 1 or 0)

Score	Meaning – criteria for scoring productive responses
2	The response conveys the speaker's feelings in a way that would be clearly understandable to an interlocutor that the speaker has just met for the first time (i.e., who does not know personal details about the speaker). The interlocutor would not need to make inferences about unknown information (a person unknown to the interlocutor). If the response is open to subjective interpretation, it is clearly culturally recognisable as interesting, boring, amazing and so on. The interlocutor would very likely recognise the cultural reference.
1	The response conveys the speaker's feelings in a way that would be somewhat understandable to an interlocutor that the speaker has just met for the first time (i.e., who does not know personal details about the speaker). If the interlocutor is required to make inferences, enough contextual information is provided for them to know generally who someone is or what something is. If the response is open to subjective interpretation, it is somewhat culturally recognisable as

	interesting, boring, amazing and so on. The interlocutor would probably understand the cultural reference but may think it odd or questionable.
0	Either the response is not understandable or does not make logical sense (see example (a)), or the result is not a simile but a literal comparison (see example (b)), or the response is too subjective because the interlocutor would not recognise the cultural reference or would likely misunderstand how the speaker feels (see example (c)) or no response is given.

Test item	Score	Example response	Example scorer justification
Q7. You recently watched a documentary on insects in the Amazon rainforest. You were amazed by how complex and organised the life of insects is. Please complete this comment to show how amazed you were by the documentary: I was so impressed by the complexity of life of those insects. It was like watching_____.	2	*miniature civilisations*	The feeling of amazement is conveyed at complexity is conveyed.
	1	*A plane fly*	Watching a plane fly might convey amazement, but it does not convey amazement at complexity and intricacy.
	0	*a) an interesting movie* *b) a tv programme* *c) Jonny Smith and his friends*[101]	a) This does not suggest anything about being impressed by complexity b) Too literal c) Too subjective
Q8. Your best friend is a very gifted athlete. She has just set a club record for the 100 metres and is focusing on training for the Olympic trials next year. You love watching her compete. Please complete this comment to show how you feel about your friend's running: My friend is one of the best sprinters in the country. When she runs at full speed, it's like watching _____.	2	*lightning*	The feeling of being impressed at speed is conveyed.
	1	*a puma*	As a big cat that lend their name to a sports brand, Pumas could be regarded as quick, however this is an odd assertion given that pumas are not known for their speed in the same way that cheetahs are.
	0	*a) fly man* *b) a sprinter*[102] *c) Me run*[103]	a) It is unclear what a fly man is and whether this would convey the appropriate feeling b) Too literal c) The interlocutor has no knowledge of the speaker's running abilities
Q9. You feel that all sandwiches from Nancy's shop are the same; boring and tasteless! Please comment to show how you feel about the sandwich's from Nancy's shop: Sandwiches from Nancy's are about as tasty as _____.	2	*cardboard*	The feeling of dissatisfaction and a bland taste is conveyed
	1	*granny's secret recipe.*	The interlocutor could infer that granny's secret recipes are bland, but the example is odd because 'secret' things tend to be desirable rather than undesirable.
	0	*a) fish and chips in the top restaurant.* *b) something disgusting*[104] *c) ones from subway.*	a) food in a top restaurant would not be understood to be bland b) Too literal c) Too subjective, the interlocutor may well misunderstand
Q10. The Smith Project was a very successful project you did in 1992. You worked so hard on the Smith Project and are very proud of what you achieved with it. Please complete this comment to show how you feel about the Smith Project: When I think of everything I've done over the years, I always come back to the Smith Project as my favourite. That project is really my _____.	2	*baby*	The feeling of pride at one's own precious creation would be conveyed
	1	*son*	The feeling of pride at one's own creation is arguably conveyed, but the specificity of 'son' makes the metaphor odd in terms of what the gender could imply, the problem that son could be grown up, in poor relations with his parent and so on.
	0	*a) piece of cake* *b) best ever piece of work*[105] *c) sun at noon*	a) The point is not to convey that the project was easy b) Too literal c) Too subjective, different people have different attitudes to the sun at noon.
Q11. You are watching a football game. Your team keep having chances to score but they miss every time. They have just missed a goal for the tenth time. Please complete this comment to show how you feel about the players in your team at the moment:	2	*an ashtray on a motorbike*	The feeling conveyed is uselessness.
	1	*ants move things to their holes but fail near the end*	The feeling of failure is arguably conveyed but the response is odd because it appears to be more an example of failure than about something not being useful

[101] Invented example
[102] Invented example
[103] Invented example
[104] Invented example
[105] Invented example

At the moment, the players are about as useful as_____.	0	*a) judge.* *b) a bad team*[106] *c) cops*	a) a judge is arguably very useful in a court of law b) Too literal c) Too subjective. The police could be perceived as useful and useless by different people with different experiences.
Q12. Your colleague Michelle is very unkind and nasty. She spreads untrue rumours about people in the office. Please complete the comment to show how you feel about Michelle: Michelle is about a nice as_____.	2	*being in the rain without an umbrella*	The feeling of unpleasantness (being soaked through) is conveyed
	1	*fox*	A fox might arguably be wily and cunning (i.e., not nice) but conversely, is also characteristically sexy (i.e., nice).
	0	*a) an angle on the earth* *b) mean woman.* *c) a talk-show actress*	a) No matter whether this is taken to mean angle or angel, it doesn't convey of regarding someone as not nice c) Too subjective
Q7. Your little niece is always jumping around and is full of energy. You adore children, especially the fact that they are full of life. Please complete this comment to show how you feel about your little niece: My niece is so energetic, she's like a little _____	2	*puppy*	The feeling of energy is conveyed
	1	*superman*	Superman is arguably energetic, but the concept of 'man' and the fact that superman is better known for strength and super powers problematises this response
	0	*a) star at night* *b) lively girl*[107] *c) brother*[108]	a) Not typically known for appearing energetic to the perceiver b) Too literal c) Too subjective, a brother might be either energetic or lazy
Q8. Your brother is very disorganised, which you hate. Please complete this comment to show how you feel about your brother's disorganisation. Let me tell you about my brother, his bedroom reminds me of a _____.	2	*a rubbish tip*	The feeling of disgust at piles of mess is conveyed
	1	*a jungle*	Jungles are arguably cluttered and messy, but a jungle might sooner convey the feeling of intrigue or heightened senses that disgust at disorganisation
	0	*a) my uncle's room that has just been rubbed.* *b) mess* *c) mixed stew*	a) Does not make sense b) Too literal c) Too subjective
Q9. Last night you went to see a choir perform at a large venue. The music and harmonies were incredible. You didn't know the choir could sing this well. Please complete the comment to show how you feel about the choir's sound: The choir I heard last night were amazing. Their sound was like_____	2	*angels rejoicing*	The feeling of amazement at beautiful sound is conveyed
	1	*owl in the nature*	arguably beautiful, but not really melodic
	0	*a) frogs in summer at evenings* *b) professional singers* *c) me singing*[109]	a) Not typically thought of as beautiful b) Too literal c) Too subjective
Q10. You are interviewing applicants for a job. The best applicant is clearly Kate. She is outstanding and much better than the others. Please complete this comment to show how you feel about Kate: We've interviewed several applicants so far, but there is one lady who is clearly _____.	2	*the front runner*	The feeling of regarding somebody as the best is conveyed
	1	*out of the crowd*	The feeling of regarding somebody as special is arguably conveyed, but it is not clear whether this is for good or bad reasons
	0	*a) our dish* *b) outstanding* *c) Jane Smith*[110]	a) Does not make sense b) Too literal c) Too subjective
Q11. You think that the decision to invest in Global Enterprises LTD would lead to disaster because people in that company are very uncooperative and difficult to work with. Please complete this comment to show how you feel about potentially working with Global Enterprises:	2	*trying to get sheep to sit together*	The feeling of frustration at a lack of cooperation is conveyed
	1	*working with a bottle of sand*	The feeling of frustration at a lack of cooperation is conveyed but the example is odd and a bottle of sand is possible confusing because it might suggest an hour glass/egg timer

[106] Invented example
[107] Invented example
[108] Invented example
[109] Invented example
[110] Invented example

Working with global enterprises would be like _____	0	*a) eating chips with rice.* *b) a disaster* *c) working with horses*	a) This does not convey frustration from a lack of cooperation b) Too literal c) Too subjective, horses can be both wild and trained
Q12. You thought that the party you attended last night was very boring. Please complete the comment to show how you feel about the party: The party was about as interesting as_____.	2	*watching paint dry*	The feeling of boredom is conveyed
	1	*wood*	Wood is in some respects a mundane material, but in other respects it is desirable (e.g., a wooden interior)
	0	*a) the joke told by a comedian.* *b) being bored*[111] *c) sleeping*	a) The interlocutor would not no which joke. b) Too literal c) Too subjective

Test 8-Idiom Extension-P (scored 2, 1 or 0)

Score	Meaning – criteria for scoring productive responses
2	The respondent has drawn on and extended the literal sense of the idiom in a way that would make sense to someone the speaker has just met for the first time (i.e., who does not know personal details about the speaker). The conceptual logic of the extended idiom as a whole is sound. 7)
1	The respondent has drawn on and extended the literal sense of the idiom in a way that would somewhat make sense to someone the speaker has just met for the first time (i.e., who does not know personal details about the speaker). The conceptual logic of the extended idiom might be questionable, but it is clear what the respondent is trying to say.
0	Either the response results in an extended idiom that is not understandable or does not make logical sense (see example (a)), extends the common, figurative sense (see example (b)) or no response is given.

Test item	Score	Example response	Example scorer justification
Q7. (Original idiom: cross that bridge when you come to it = wait to deal with a problem only if or when it happens) Extended idiom: Let's cross that bridge when we come to it. Although, since the decision seems likely, let's _____. Please extend the idiom:	2	*start figuring out how to cross safely!*	The response extends the idea of a bridge and has conveyed the meaning 'let's start to think about how to deal with the problem'
	1	*jump over it*	The response extends the idea of proceeding over the bridge but it is unclear why the bridge would need to be jumped over
	0	*a) be prepared for the battlefield* *b) deal with it now*	a) A different metaphorical domain (battle) has been used b) Too literal
Q8. (Original idiom: beat around the bush = to avoid answering a question of making a clear point when talking) Extended idiom: He beat around the bush for so long that _____! Please extend the idiom:	2	*he got dizzy and fell over!*	The response extends the idea of beating around the bush and has successfully emphasised the person's indecision
	1	*he has totally passed the bush*	The response extends the idea of beating around the bush but it is not logical that a person beating around a bush would then pass it
	0	*a) no one punch on the point* *b) I cannot get the point what he is talking about*	a) It is not clear what this means b) Too literal
Q9. (Original idiom =to take the cake = to be outstanding either in a very good or a very bad way) Extended idiom: His comment really took the cake. In fact it didn't just take the cake, it _____. Please extend the idiom:	2	*took the whole picnic!*	The response extends the idea of taking a piece of cake and has successfully emphasised the impact of the comment
	1	*has icing on it as well*	The response extends the idea of taking the cake, but it is unclear why the cake taker (i.e., the comment personified) now has icing on it (unless some icing got transferred in the process).
	0	*a) also take the fat* *b) is extremely good*	a) Fat is not a plausible component of a meal involving cake b) Too literal
Q10. (Original idiom: to make a mountain out of a molehill = to make a small problem seem very dramatic or important)	2	*he started operating hiking excursions!*	The response extends the idea of the growing molehill/mountain and has successfully emphasised the fuss made about the problem.

[111] Invented example

Extended idiom: He made such a mountain out of a molehill that _____. Please extend the idiom:	1	*the molehill was blocked by the mountain*	The response extends the idea of idea of the growing molehill/mountain but it is unclear why what was at first a molehill becoming a mountain is now two separate entities
	0	*a) everyone give him a cold shoulder* *b) things became so dramatic*	a) This constitutes a different metaphor and unwarranted reference b) Too literal
Q11. (Original idiom: sitting on the fence = not making a decision about something) Extended idiom: He seems to be sitting on the fence about it. In fact, he's been sitting on the fence so much that _____. Please extend the idiom:	2	*his wife has brought him a newspaper and a glass of lemonade!*	The response extends and successfully emphasises the idea of a man sitting on a fence for a long time.
	1	*he cannot get off the fence*	The response extends the idea of a man sitting on a fence but the absence of detail concerning why the man cannot get off the fence makes this odd, as does the repetition of the word fence.
	0	*a) we do not know when will he build the fence* *b) he might be struggled with this decision*	a) this is not logical, the fence already exists b) Too literal
Q12. (Original idiom: to fell head over heels in love = to be very much in love) Extended idiom: She fell so head over heels in love that _____. Please extend the idiom:	2	*she rolled all the way down the hill!*	The response extends the idea of a girl falling to the extent that she rolls all the way down a hill, i.e., falls very much in love
	1	*her head is as low as her heels*	The response extends the idea of the girl falling, but it the fact that her head is now as low as her heals is an odd thing to point out since the rest of her body would surely be as low as her heals.
	0	*a) being crazy about his boyfriend* *b) she loves him very much*	a) This is not an extended idiom involving two men b) Too literal
Q7. (Original idiom: to get a taste of your own medicine = to receive the same unpleasant experience that you yourself have given to someone else.) Extended idiom: He got such a taste of his own medicine that _____. Please extend the idiom:	2	*he exceeded the recommended daily dosage!*	The response extends the idea of taking one's own medicine in large quantities and successfully emphasises the original idiom
	1	*it's really disgusting*	The response extends the idea of The response extends the idea of taking one's own medicine but is odd because 'it's really disgusting' is not a consequence of taking large quantities of something, just tasting it in the first place.
	0	*a) he could healed by others.* *b) he deserves it*	a) It does not logically follow that the person could be healed by other people b) Too literal
Q8. (Original idiom: to be stuck between a rock and a hard place = to be in a very difficult situation) Extended idiom: It was a difficult decision. We were so stuck between a rock and a hard place that _____. Please extend the idiom:	2	*our feet were beginning to resemble fossils!*	The response extends the idea of feet being stuck fast in the ground and emphasises the original idiom by way of a possible consequence of being stuck for a (very) long time.
	1	*we could not move forward.*	The response extends the idea of being stuck fast in ground but there are conceptual problems because the person would not be aiming to move forward, but rather to become unstuck
	0	*a) we really get lost* *b) WE ARE IN A Dilema*	a) This implies that the people are currently moving, which they cannot be b) Too literal
Q9. (Original idiom: the ball is in your court = it's your turn to respond or take action) Extended idiom: After her email the ball is in my court. But the problem is _____. Please extend the idiom:	2	*I didn't want to play anymore!*	The response extends the idea of playing ball and implies that the person does not want to exchange emails anymore
	1	*the ball sticked to my grass!*	The response extends the idea of playing ball but it is unclear why the ball might get stuck to grass
	0	*a) I do not know who did it*	a) It is unclear what this means b) Too literal

		b) I lost the key to the email box.	
Q10. (Original idiom: to hit the nail on the head = to identify exactly what is causing an issue or problem) Extended idiom: When he said that, he became the first person to really put the problem into words. And he hit the nail on the head so hard that_____. Please extend the idiom:	2	*we all felt it go through the wood!*	The response extends the idea of a nail being hit and suggests a consequence that can be understood as meaning the remark was very poignant
	1	*Almost the whole head has been hit.*	The response extends the idea of a nail being hit and a possible (but odd) consequence, but it unclear as to why the whole head was not hit
	0	*a) every catch the important point* *b) he found out the problem very quickly*	a) It is unclear what this means b) Too literal
Q11. (Break a leg! = do your best!) Extended idiom: Don't worry, your performance will be great! Just go out and break a leg. In fact, go out and_____! Please extend the idiom:	2	*come back with crutches!*	The response extends the idea of breaking a leg and emphasises the original idiom in a way that can be understood as 'go and do your very very best!'
	1	*break it at the worst ever*	The response extends the idea of breaking a leg but it is not fully clear what 'at the worst' could mean
	0	*a) get some fresh air* *b) try your best*	a) Not logically involved with breaking a leg b) Too literal
Q12. (Original idiom: it's raining cats and dogs = it's raining a lot) Extended idiom: It's been raining cats and dogs for so long that _____. Please extend the idiom:	2	*we've had to call the stray animal collection agency!*	The response extends the idea of raining cats and dogs and provides a logical consequence that could be understood as meaning 'it's been raining heavily for a very long time'
	1	*it turns to be a zoo*	The response extends the idea of raining cats and dogs but a zoo would not be the logical result (zoo's do not primarily contain domestic animals)
	0	*a) we cannot find a place to land on* *b) the rain is too heavy to go out.*	a) This is illogical, 'we' are not falling b) Too literal

Test 9-Metaphor Continuation-P (scored 2, 1 or 0)

Score	Meaning – criteria for scoring productive responses
2	The response keeps the code going or is written in code (i.e., using metaphor) in a way that is clearly understandable in conjunction with the preceding dialogue. The conceptual logic of the utterance is sound.
1	The response keeps the code going or is written in code (i.e., using metaphor) in a way that is somewhat understandable in conjunction with the preceding dialogue. The conceptual logic of the utterance might be questionable, but it is clear what the respondent is trying to say.
0	Either the response results in an utterance that is not understandable in conjunction with the preceding dialogue or does not make logical sense (see example (a)), or the response is not in code in the sense that it is literal (see example (b)) or no response is given.

Test item	Score	Example response	Example scorer justification
Q7 Your mum: Jack, the machine called in earlier! You: Haha, I know we joke about it, but it's really true; he is a machine! You can always see that he is_____. **Please write responses in 'code' that keep the conversation with your mum going:**	2	*he is switched on and in motion*	This continues the code of the brother as a machine and can be understood to mean 'he is active'
	1	*working*	A machine can be described a working, but so can a human, so this receives 1
	0	*a) earlier than machine.* *b) calculating his next move.*	a) This does not make sense b) Too literal
Q8 Your mum: You'll never believe it, he steamed over to the house in search of midday fuel, again!	2	*go back to his own petrol station*	This continues the code and can be understood to mean 'go back to his own house for lunch'

You: That sounds about right! Even though he left home several years ago, he still comes here for refuel. Why didn't you just tell him _____!? **Please write responses in 'code' that keep the conversation with your mum going:**	1	*machine needs maintenance.*	This continues the code but it is not exactly the point that the machine needs maintenance now
	0	*a) the truth of house* *b) to let him in*	a) this does not make sense b) Too literal
Q9 Your mum: well, it was quite nice to feel like the mechanic again, or at least the petrol station attendant! He actually seemed a bit conked out You: Really, well, I'm sure that after receiving his refuelling and a bit of home mechanics, he's now _____! **Please write responses in 'code' that keep the conversation with your mum going:**	2	*burning rubber again!*	This continues the code and can be understood to mean 'being active again'
	1	*quite full*	This continues the code but it is dubious as to whether it means literally full of food or full of fuel
	0	*a) very sad* *b) refreshed*	a) This is not logical b) Too literal
Q10 Peter: Have you heard, the wizard has done his magic again? I mean the secret magic award You: oh yes, I heard Mr magic is due to be _____. **Please write responses in 'code' that keep the conversation with Peter going:**	2	*formally recognised for his new and inspiring spells*	This continues the code and can be understood to mean *receive a pay raise, promotion, and so on*.
	1	*getting the magic award*	This continues the code but repeats 'award'
	0	*a) out of magic* *b) popular*	a) This is not logical, he is at the 'height of his powers' b) Too literal
Q11 Peter: Yes, that's right, his spells have been creating quite a positive stir in the kingdom You: Which spell in particular? Will the magic circle commend him for_____? **Please write responses in 'code' that keep the conversation with Peter going:**	2	*putting such a spell on our clients*	This continues the code and can be understood to mean 'impressing the clients with something'
	1	*stirring the kingdom heavier*	This continues the code but it is unclear what heavier means, perhaps with more vigour?
	0	*a) death* *b) his hard woring*	a) Not logical b) Too literal
Q12 Peter: I think his main magical achievement was something like that. But he's really all-round enchanting; he's simply been running our show for a long time You: I agree, I'm completely _____. **Please write responses in 'code' that keep the conversation with Peter going:**	2	*spellbound*	This continues the code and can be understood to mean 'amazed'
	1	*attacked by his spells*	This continues the code but it is unclear what attacked could refer to (e.g., jealous of?)
	0	*a) board with that.* *b) not doing my work*	a) assuming this means bored it does not fit with the rest of the dialogue b) Too literal and not logical
Q7 John: Hi, it's my lunch break… On my laptop so need to write covertly in case anyone walks past and glances at the screen :)…you remember 'operation C'? You: Hi John, haha yes I remember. How _____? **Please write responses in 'code' that keep the conversation with John going:**	2	*is the operation unfolding*	This continues the code and can be understood to mean 'are things going with your new job application'
	1	*about shortcut key S*	This continues the code but is very ambiguous, though it could be the suggestion to introduce a new code word (e.g., s for success?)
	0	*a) please I was with the outcome* *b) is it going*	a) this does not fit with the dialogue b) Too literal
Q8 John: Well I've been in to assess the lay of the land, me and some rival agents met with a strict panel of drill Sergeants if you know what I mean :) It seems they've chosen their James Bond, yours truly ;) You: Wow, that's excellent news! So you are saying_____? **Please write responses in 'code' that keep the conversation with John going:**	2	*you'll be allied to a different government soon*	This continues the code and can be understood to mean 'you'll be switching employers soon'
	1	*that you will be the James Bond for him?*	This continues the code but it is unclear who him is (the panel of drill sergeants is plural)
	0	*a) the one in the town* *b) you will leave for a new job?*	a) this does not make sense b) Too literal
Q9 John: That's right. To be honest, I'm a bit worried about how to switch over from my current operation if you catch my drift :) The crew and captain will not be very pleased that I'm jumping ship! You: Well think of it like this: _____. Don't worry, it'll be fine. **Please write responses in 'code' that keep the conversation with John going:**	2	*every operation comes to an end*	This continues the code and can be understood to mean 'every period of tenure has to end'
	1	*you dive like a fish into the water*	This continues the code but it is unclear, though it perhaps means go for it
	0	*a) throwing your hat into the river* *b) it will be better for u*	a) This is completely unclear b) Too literal
Q10 Mary: Hey! It's Mary, I've got great news, that I'll tell you in code :)…you know I've been really hungry these past few weeks? Well today the doctor confirmed that I'm eating for two now ;)	2	*you've got a bun in the oven*	This continues the code and can be understood to mean 'you're pregnant'
	1	*you have carried another young fellow in your body?*	This continues the code but the tense makes this odd

You: Hi Mary, Wow! So you're telling me that _____? **Please write responses in 'code' that keep the conversation with Mary going:**	0	*a) you have a sick* *b) you are pregnant with two*	a) This is perhaps an incorrect interpretation of the situation (i.e., visiting the doctor) b) Too literal
Q11 Mary: Yep that's right :D The stork will be paying me a visit around March 15th next year :) You: Great! That's fantastic news! What about gender? Will you_____? **Please write responses in 'code' that keep the conversation with Mary going:**	2	*be buying pink or blue*	This continues the code and can be understood to mean 'buying clothes for a girl or boy'
	1	*having the male stork or female one?*	This continues the code but it is not the gender of the stork that is important (though this is understandable)
	0	*a) the lonely sheep in the group?* *b) have a boy or girl*	a) This is not understandable b) Too literal
Q12 Mary: I don't know yet, it's far too early, but I'll be announcing it formally in a couple of weeks. You: That's wonderful, I'm so glad to hear that once again, you'll be _____:) **Please write responses in 'code' that keep the conversation with Mary going:**	2	*extending the family*	This continues the code and can be understood to mean 'having another baby'
	1	*carrying a ball?*	This continues the code but if the ball represents a baby this would be an odd metaphor
	0	*a) the brightest star in the sky!* *b) a mother*	a) this is very unclear, and perhaps too vague (the woman does not excel at something as the metaphor would suggest) b) Too literal

Appendix B MC Test Battery (version 1) as seen by group 1 participants

[NOTE: TESTS WERE PRESENTED TO PARTICIPANTS AS 'SECTIONS']

Introduction

These are questions that I have designed for my research and are not linked to your work or studies (so don't worry!). The questions all require multiple-choice selection or short answers; there are no long essay questions. Here is some useful information about answering the questions:

- Please answer all the questions as best you can. You must provide an answer in order to proceed to the next question; if you are not sure, just guess or write '?' and move on. There are 9 sections, so plenty of chances to write good answers elsewhere! :)
- For some of the questions, there is no one right answer, so do not worry about getting everything correct.
- Please work at a good pace, do not spend too long on any one answer.
- Please take breaks when you need to, but do aim to finish the test on the same day that you start it.
- The test saves your data as you go; here is no need to click save at any point. When you get to the end of the test you will be notified.
- If the screen crashes, click the link to the test in the email that I sent you, this will take you back to where you were before the crash.
- Once you have chosen your answer to a question, click 'next' at the bottom. You will not be able to go back and change your answer once you have done this.

Please click the purple button (>>) to proceed to the first section.

Section 1: Part A

Instructions

In part A you will be choosing the right word (from a multiple-choice) to complete some two word phrasal verbs. There is a clue in brackets. e.g., We're just waiting for one more person to **turn** _____(arrive), then we'll start the meeting.

a) over
b) out
c) up
d) down

The correct answer is c) up. Here, the clue is "arrive". The answer you provide will need to give the same meaning as the clue (turn up = "arrive"). **Note:** sometimes, the sentence actually makes sense as it is, but ignore this...you will always need to choose one of the four options to complete the two word phrasal verbs.

Questions

Q1.1. Business has been very poor but we expect it to **pick** _____ (improve) again before Christmas.

- on
- off
- away
- up

Q1.2. The police officer who spoke to us wanted to **take** _____ (record) all of our details.

- on
- in
- up
- down

Q1.3. If this information **gets** _____ (becomes public), it will be the end of her career as a politician!

- away
- on
- over

- out

Q1.4. We asked all teachers to **give** _____ (distribute) a general reminder to students.

- out
- off
- away
- over

Q1.5. The tickets are too expensive; people might be **put** _____ (discouraged) from attending.

- out
- down
- off
- away

Q1.6. I'll try to **get** _____ (do) an hour of reading before dinner.

- out
- on
- in
- with

Q1.7. Do we need to **put** _____ (record) any other names on the list of invites?

- down
- in
- across
- on

Q1.8. How many days will I need to **take** _____ (be absent) from work after my operation?

- out
- up
- off
- on

Q1.9. Schools usually **break** _____ (stop) for summer in the middle of July.

- away
- down
- off
- up

Q1.10. There's been an accident. We're still waiting for more news to **come** _____ (arrive).

- up
- in
- on
- over

Section 1: Part B

Instructions

In part B you will type in your own answers. E.g., We're just waiting for one more person to **turn** _____(arrive), then we'll start the meeting. You should type ''up'…*We're just waiting for one more person to **turn up** (arrive), then we'll start the meeting.* **Note:** you will always need to type in a ''particle'' (e.g., up, on, in, under) to complete the two word phrasal verbs. Sometimes the sentence makes sense even before anything is added, but please always type in a ''particle'' (e.g., up, on, in, under).

Questions

Q1.11. He spoke really quickly. Did you manage to **get** _____ (record) everything he said?

Q1.12. Don't let the quality of your work **go** _____ (decrease)!

Q1.13. One of the boxers was much stronger, so we knew who would **come** _____ (emerge) worse.

Q1.14. I'm not asking you to **put** _____ (contribute) too much time, just one or two hours a week.

Q1.15. I want to **move** _____ (get promoted) to a more senior position in my company next year.

Q1.16. They'll probably **get** _____ (escape) with a warning this time, but it was a very stupid thing to do.

Q1.17. Just park here and unload; you won't **hold** _____ (block) any traffic at this time of night.

Q1.18. With this new job, I can **bring** _____ (earn) enough money to pay my daughter's tuition fees.

Q1.19. We don't want the campfire to **go** _____ (become extinguished), so let's find more wood.

Q1.20. I'll help you **pick** _____ (choose) a new dress from the selection in the store. You're going to look beautiful at the formal dinner!

Section 2: Part A

Instructions

In this section, you will:

a) explain the meaning of an expression
b) answer a multiple-choice question about the meaning of the expression.

E.g., a) **Sentence**: It was a solid argument.

Example explanation: *This means...that the argument was good, strong and valid*

This is a good answer; the meaning has been explained correctly (green text) using a full sentence.

Sentence: It was a solid argument.

Question: *Which of the following options is best for helping us understand the meaning of this sentence?*

a) The idea of the cost of buildings in the 21st century
b) The idea that the argument cannot easily be destroyed
c) The idea of liquids that have become solid
d) The idea that the argument was difficult to understand

The correct answer is b). The other options contain information and ideas that are not really relevant to helping us understand "it was a solid argument".

Questions

Q2.1. The news lifted her spirits

a) this means...

b) which of the following options is the best for helping us understand the meaning of this sentence:

- The idea of strength involved in lifting
- The idea of breathing air into the chest
- The idea of the sound of straining as something is lifted
- The idea of feeling lighter in the chest

Q2.2. She treated us in a very cold way.

a) this means…

b) which of the following options is the best for helping us understand the meaning of this sentence:

- The idea of temperature on a thermometer
- The idea of not wanting to have contact with cold things
- The idea of the appearance of ice and snow
- The idea of receiving cold food

Q2.3. They will want to get married sometime in the distant future.

a) this means…

b) which of the following options is the best for helping us understand the meaning of this sentence:

- The idea that it is expensive to travel long distances
- The idea that people often get tired when they travel
- The idea of travelling towards a destination
- The idea of needing to buy things for a journey

Q2.4. He has a fiery temper.

a) this means…

b) which of the following options is the best for helping us understand the meaning of this sentence:

- The idea of using fire for cooking
- The idea that fire can be frightening
- The idea that burning things smell
- The idea that fire requires oxygen to burn

Q2.5. The conscience is man's compass.

a) this means…

b) which of the following options is the best for helping us understand the meaning of this sentence:

- The idea that a compass can be broken
- The idea of the price of a compass
- The idea that west is good and east is bad
- The idea of a true and good direction

Q2.6. TV is chewing gum for the eyes.

a) this means…

b) which of the following options is the best for helping us understand the meaning of this sentence:

- The idea that chewing gum does not have much nutritional value
- The idea that chewing gum is colourful
- The idea of different brands of chewing gum
- The idea of the shape of a piece of chewing gum

Section 2: Part B

Instructions

In part B you will answer multiple-choice questions. Each sentence was said by a comedian. The option that you choose should be the best one for making the sentence **funny** or **witty**. For example:

Sentence: John has a big problem...

a) he's become quite unhappy!
b) he's very angry!
c) he's in debt!
d) he's overweight!

The best answer is **d) he's overweight**! This is because it makes a joke based on two meanings of 'big'. If we read 'John has a big problem...' and stop reading there, we understand that 'big' means significant or great. If the full sentence is 'John has a big problem...he's overweight', then we re-understand 'big' to mean fat, physically large.

The other choices, a), b), and c) do not make the sentence **funny** or **witty** in this way.

Questions

Q2.7. In the mirror I looked like a million dollars, _____

Please choose the best answer for making the sentence funny or witty:

- green and wrinkled!
- wonderful!
- a bundle of paper bills!
- sick and old!

Q2.8. When everything's coming your way, _____

Please choose the best answer for making the sentence funny or witty:

- you're in the wrong lane!
- life is great!
- you could be involved in a car crash!
- life is a disaster!

Q2.9. No person goes before their time, _____

Please choose the best answer for making the sentence funny or witty:

- unless the boss leaves early!
- only when they are due to leave this world!
- they must wait until the end of the working day!
- everyone leaves this world early!

Q2.10. My wife's currently carrying our first child, _____

Please choose the best answer for making the sentence funny or witty:

- he's eight years old the lazy little thing!
- she's pregnant!
- he's such a little baby, and so light to hold!
- his brother can't wait for him to be born!

Q2.11. The only thing moving about this actor's performance _____

Please choose the best answer for making the sentence funny or witty:

- was his wig!
- was his incredible acting!
- was his body!
- was his terrible singing!

Q2.12. The young fighter had a hungry look, _____

Please choose the best answer for making the sentence funny or witty:

- the kind you get from not eating for a while!
- the kind you get when you really want to win!
- the kind that expresses your hunger for food!

- the kind that means you are ready to quit!

Q2.13. My local police chief does a talk on drugs, _____

Please choose the best answer for making the sentence funny or witty:

- you can't understand half of it!
- it was well-structured but a bit boring!
- he stepped off them at the end!
- he completely covered the topic!

Q2.14. When I found out my toaster was not waterproof_____

Please choose the best answer for making the sentence funny or witty:

- I was shocked!
- I was surprised!
- I was electrocuted!
- I was physically traumatised!

Q2.15. Never trust an atom, _____

Please choose the best answer for making the sentence funny or witty:

- they make up everything!
- they compensate for everything!
- they constitute everything!
- they apply cosmetics to everything!

Q2.16. If I were two faced, _____

Please choose the best answer for making the sentence funny or witty:

- I would not be wearing this one!
- I would talk badly about people without them knowing!
- I would seek medical help to get one removed!
- I would look sad!

Q2.17. My friends and I put together a performance on puns; it was basically just_____

Please choose the best answer for making the sentence funny or witty:

- a play on words!
- a manipulation of language!
- a show about sentences!
- a fun time with grammar!

Q2.18. True friends_____

Please choose the best answer for making the sentence funny or witty:

- stab you in the front!
- stab you in the back!
- stab you in the heart!
- stab you in the little finger!

Section 3

Instructions

In this section you will rate the acceptability of some expressions. English native speakers often use expressions which mix ideas and concepts in what seems like quite a strange way. For example, they talk about people as if they were plants or fruit, e.g.,

1. He's so **rotten**! (= he's mean or cruel)
2. She's really **blossomed** into an attractive young lady! (= she's become very attractive)

Sentences 1 and 2 are both perfectly acceptable English expressions. Other expressions connected to this idea are not possible, e.g.,

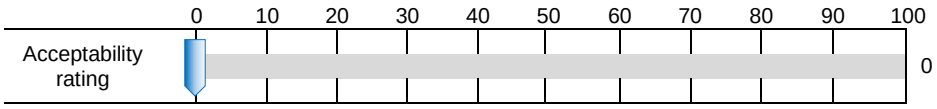
3. We **potted** her
4. He **photosynthesised** last week

Sentences 3 and 4 use the same idea (people are plants) and are grammatically correct, but they sound very strange and are difficult to understand. For each of the following sentences, please rate the acceptability of the expression in bold by dragging the slide. An acceptable expression is one that an English native speaker might use in the context of the sentence (to rate 0, you will still need to click the slide).

Questions

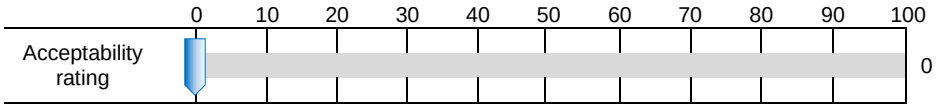
Q3.1. His **blood began to boil** and he started shouting.

0= not acceptable in English, 100 = perfectly acceptable in English.



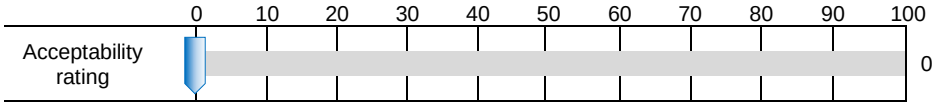
Q3.2. He **slipped into** a depression

0= not acceptable in English, 100 = perfectly acceptable in English.



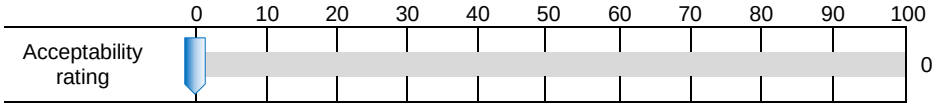
Q3.3. His body **went** fat after a few years

0= not acceptable in English, 100 = perfectly acceptable in English.



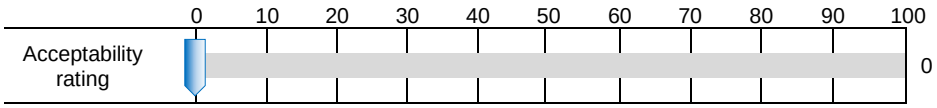
Q3.4. The whole theory **fell apart**

0= not acceptable in English, 100 = perfectly acceptable in English.



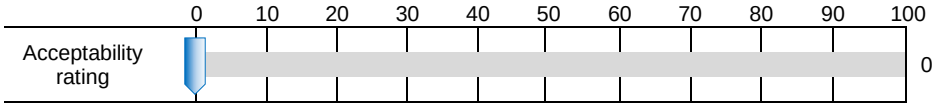
Q3.5. The project is **going ahead** as planned

0= not acceptable in English, 100 = perfectly acceptable in English.



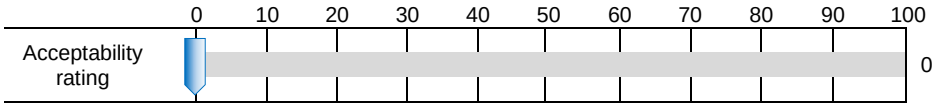
Q3.6. He **couldn't bottle his anger up** anymore so he started shouting

0= not acceptable in English, 100 = perfectly acceptable in English.



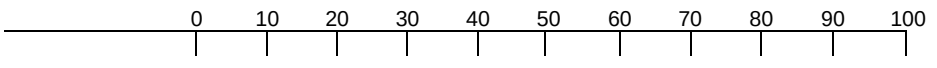
Q3.7. It was an **attractive** proposal

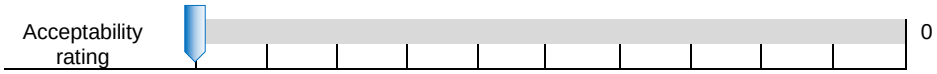
0= not acceptable in English, 100 = perfectly acceptable in English.



Q3.8. The idea **holds up** in principle

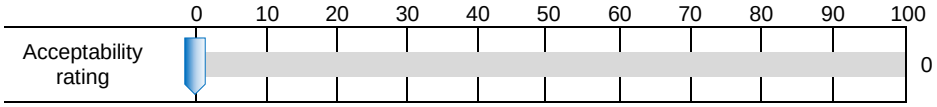
0= not acceptable in English, 100 = perfectly acceptable in English.





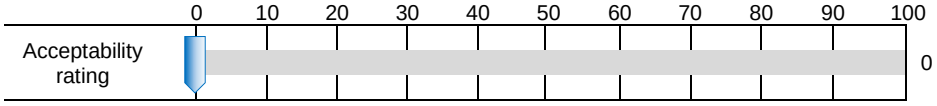
Q3.9. To her the drunken man was **repulsive**

0= not acceptable in English, 100 = perfectly acceptable in English.



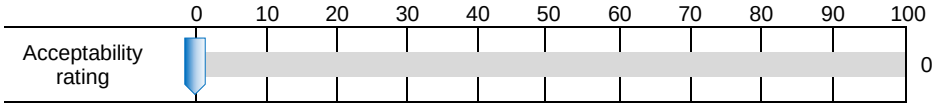
Q3.10. The theory was **the colour of brick**

0= not acceptable in English, 100 = perfectly acceptable in English.



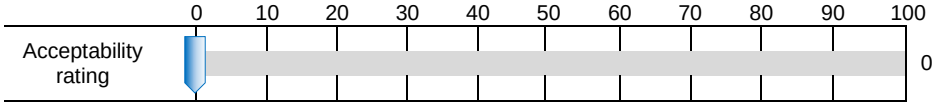
Q3.11. There was a lot of **electricity between** the dog and ball

0= not acceptable in English, 100 = perfectly acceptable in English.



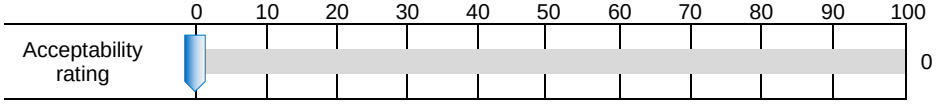
Q3.12. Her hair had almost **arrived at** being grey

0= not acceptable in English, 100 = perfectly acceptable in English.



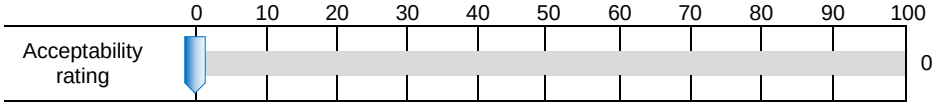
Q3.13. We **entered the front door** of the plan

0= not acceptable in English, 100 = perfectly acceptable in English.



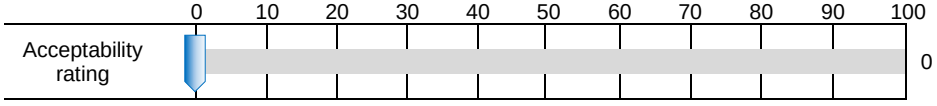
Q3.14. He **bubbled** as he began shouting

0= not acceptable in English, 100 = perfectly acceptable in English.



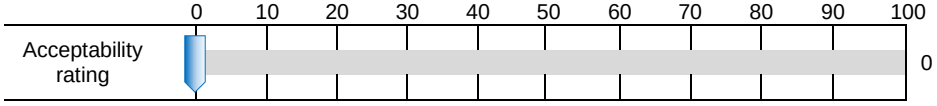
Q3.15. Their similarities **jerked** them together

0= not acceptable in English, 100 = perfectly acceptable in English.



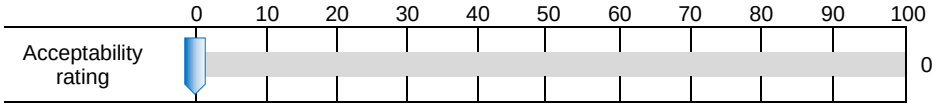
Q3.16. She **turned orange** as she started shouting at him

0= not acceptable in English, 100 = perfectly acceptable in English.



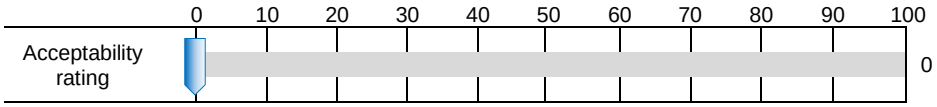
Q3.17. He **freshened** his ideas

0= not acceptable in English, 100 = perfectly acceptable in English.



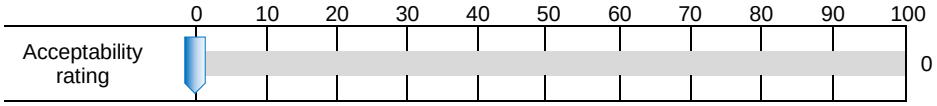
Q3.18. He told a **white** lie

0= not acceptable in English, 100 = perfectly acceptable in English.



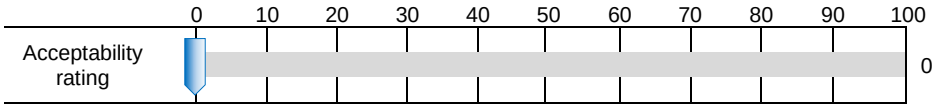
Q3.19. She made a **firm** proposal to the client

0= not acceptable in English, 100 = perfectly acceptable in English.



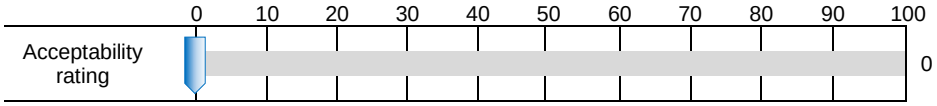
Q3.20. He tried **to pull the wool over my eyes**

0= not acceptable in English, 100 = perfectly acceptable in English.



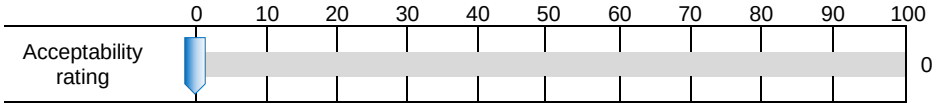
Q3.21. He never has time **to shoot the breeze**

0= not acceptable in English, 100 = perfectly acceptable in English.



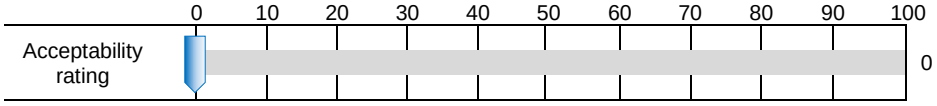
Q3.22. He has a **killer** headache

0= not acceptable in English, 100 = perfectly acceptable in English.



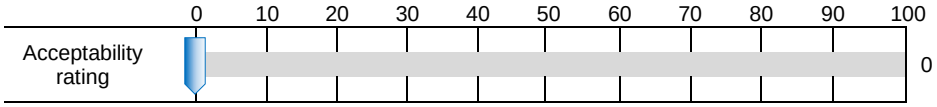
Q3.23. We solved the **teased out** problem very easily

0= not acceptable in English, 100 = perfectly acceptable in English.



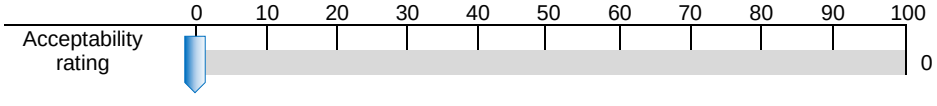
Q3.24. I **picked up** a job last week

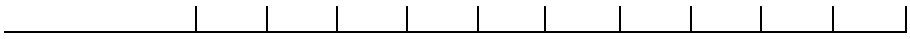
0= not acceptable in English, 100 = perfectly acceptable in English.



Q3.25. The comment **blunts**

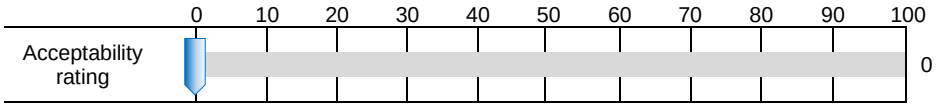
0= not acceptable in English, 100 = perfectly acceptable in English.





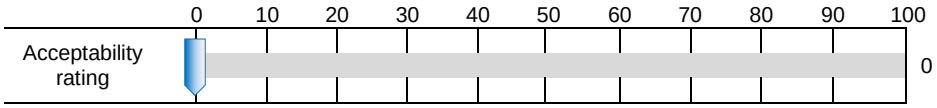
Q3.26. We asked for **a called day** at 6pm

0= not acceptable in English, 100 = perfectly acceptable in English.



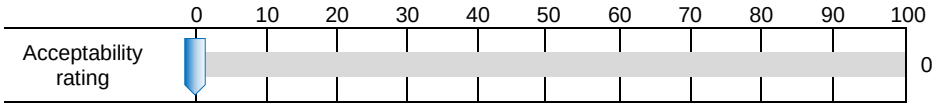
Q3.27. I will give you **a show of the ropes** tomorrow

0= not acceptable in English, 100 = perfectly acceptable in English.



Q3.28. The team are trained to makes calls **coldly**; customers never expect their calls!

0= not acceptable in English, 100 = perfectly acceptable in English.



Section 4: Part A

Instructions

In part A you will rate the suitability of options for filling a gap. Each sentence in this section is an incomplete analogy (an analogy is a statement that helps us understand one thing by comparing it with another thing). For each sentence, please rate each of the four options according to how well they complete the analogy. E.g.,

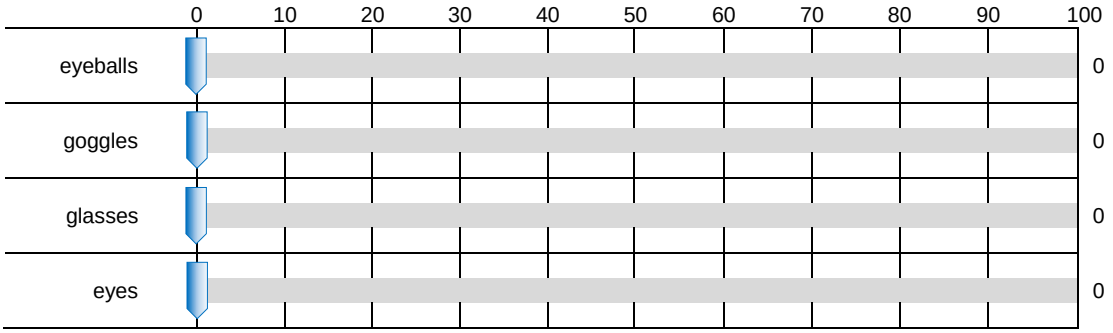
Mark is the _____ of our organisation; he is in charge and rules the place!

a) king (< this would get 90/100, a king rules, and so does Mark, it forms a good analogy)
b) citizen (< this would get 20/100. A citizen does not rule, so this is not very helpful)
c) queen (< this would get 10/100. A queen is female; Mark is male)
d) jester (< this would get 20/100. A jester is usually male (like Mark) but is a low ranking member of the court, he doesn't rule)

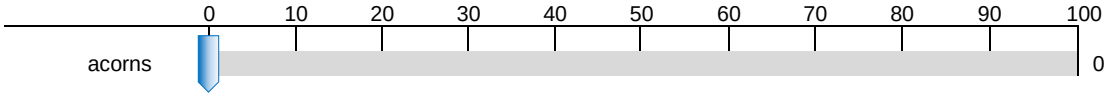
Please rate all four options. To give a rating of 0, please click on the option (you will need to do this to proceed to the next question).

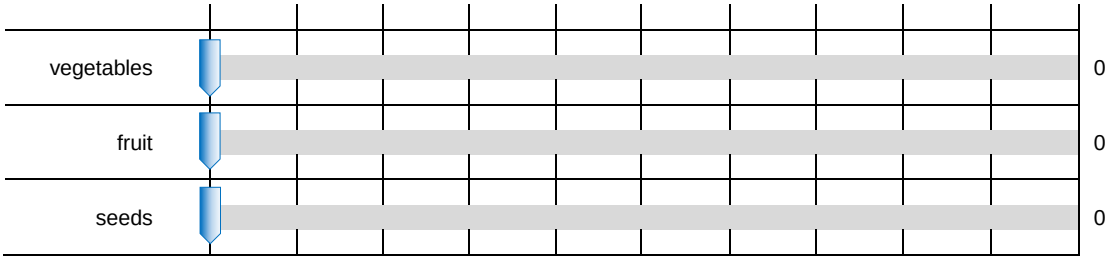
Questions

Q4.1. The CCTV cameras are the _____of the building.

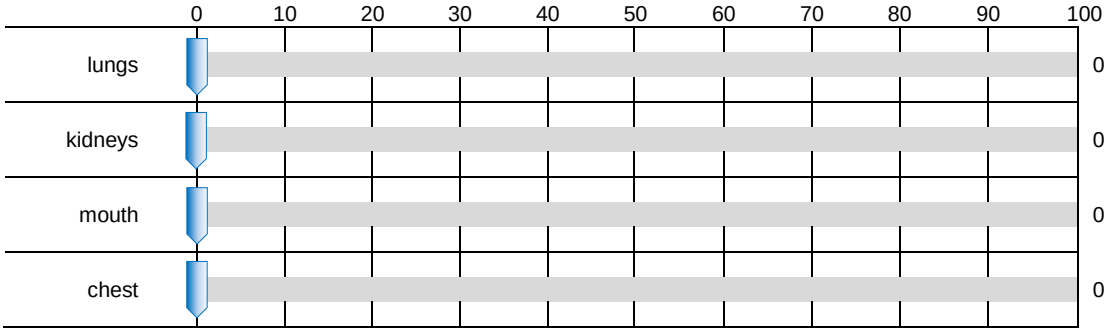


Q4.2. New products at the end of a long production process are the _____of large companies.

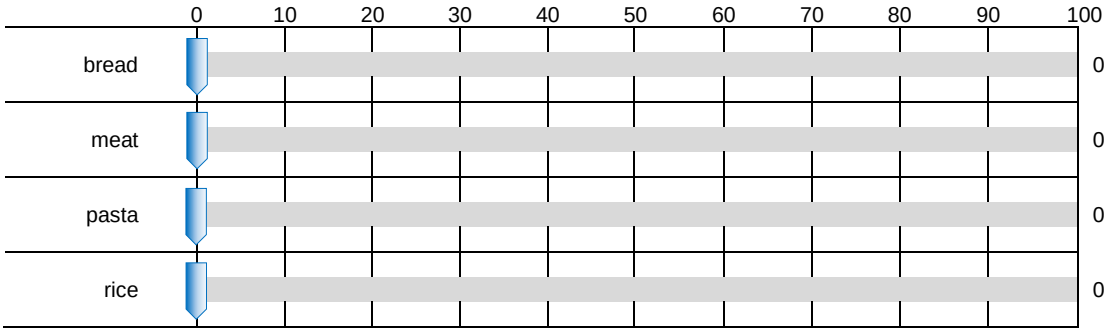




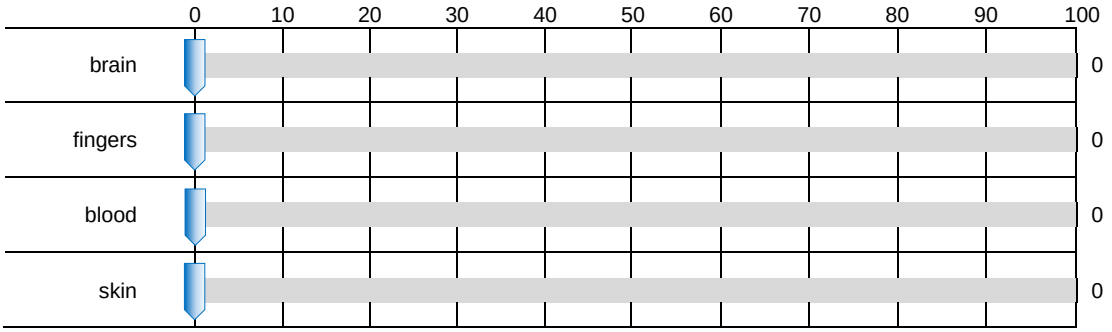
Q4.3. This park is the _____of our city.



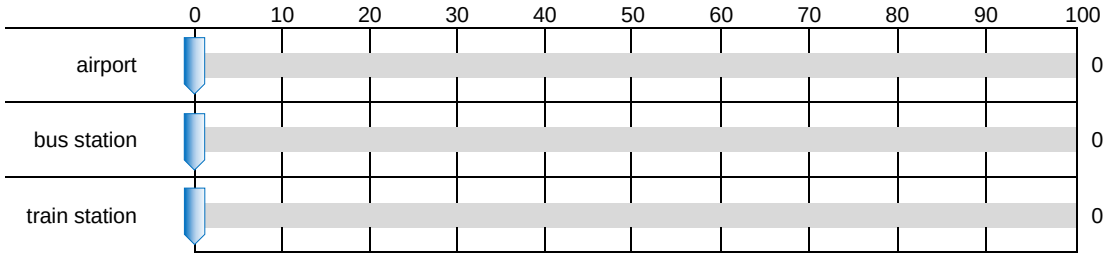
Q4.4. The main argument is the _____of the essay.

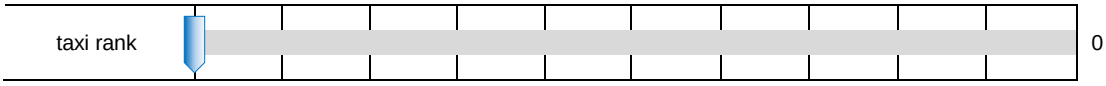


Q4.5. The company's internal mail team are the _____of the organisation.



Q4.6. The bee hive is the _____ of the animal kingdom.





Section 4: Part B

Instructions

In Part B, you will type in your own answers.

Questions

Q4.7. The sales team are the _____of the organisation.

Q4.8. Alcohol is the _____of the drunk person.

Q4.9. The outside walls are the _____of a building.

Q4.10. Killer whales are the _____of the sea.

Q4.11. Volcanoes are the _____of the earth.

Q4.12. Chemical elements are the _____of life.

Section 5: Part A

Instructions

In part A you will answer multiple-choice questions.

At the end of a conversation, we often use an expression to summarise the main point, specify some overall advice, and/or let the other speaker know that we would like to change topic. **Please choose the best expression to finish the conversations.**

Questions

Q5.1. Speaker A: Did I tell you that Sarah and I broke up last week?
Speaker B: No! Oh that's so sad, how come?
Speaker A: We just weren't right for each other. I'm so down; I just don't feel like I'll ever meet the right person.
Speaker B: I'm sure you will, I know Sarah was great but don't worry, you know what they say, _____

Please choose the best expression to finish the conversation:

- give a man a fish, and you feed him for a day; show him how to catch fish, and you feed him for a lifetime!
- there's plenty more fish in the sea!
- telling a teenager the facts of life is like giving a fish a bath!
- no human being, however great or powerful, was ever so free as a fish!

Q5.2. **Speaker A:** We went to a small village in France last month.
Speaker B: Oh that's great, what did you do?
Speaker A: Well, we really tried to enjoy the French culture and fit in with the locals. We drank fresh coffee and read a newspaper in the mornings, then ate lunch with wine, and in the evening walked the streets listening to live music playing in the restaurants. We really started to feel French!
Speaker B: Great, well you know what they say, _____

Please choose the best expression to finish the conversation:

- Rome wasn't built in a day!
- Nero found Rome built of bricks and left it clothed in marble!
- even the Romans couldn't conquer the blue skies!
- when in Rome do as the Romans do!

Q5.3. **Speaker A:** I had an interesting dilemma the other day, my boss asked me to prepare a report over the weekend, to be ready for Monday morning. Unfortunately, I had a lot of stress with the furniture removal people on Saturday and I just forgot to do the report. It's not a good excuse but it's the truth.
Speaker B: So what did you tell your boss?
Speaker A: In the end I decided not to lie and that it was better to tell the truth and apologise.
Speaker B: I agree, you know what they say, _____

Please choose the best expression to finish the conversation:

- the truth is hard to come by!
- honesty's the best policy!
- truth is stranger than fiction!
- better to tell some home truths!

Q5.4. **Speaker A:** It was so difficult to find funding for my studies. I applied to seven different funding councils, all of them rejected me. I then looked at loan options and part time work. It was tough but I was so determined that I would find funding and start my studies.
Speaker B: So did you have any success?
Speaker A: Yes, I managed to get funding from the company I currently work for, they have a scheme for employees looking to continue their education.
Speaker B: That's great, you know what they say, _____

Please choose the best expression to finish the conversation:

- the spirit is willing but the flesh is weak!
- never spur a willing horse!
- where there's a will there's a way!
- you can lead a horse to water but you can't make it drink!

Q5.5. **Speaker A:** I'm so embarrassed.
Speaker B: Why, what happened?
Speaker A: Yesterday, I accidentally broke my colleague Peter's coffee mug in our office kitchen
Speaker B: Did he get angry?
Speaker A: Well, he was away on business yesterday, but he's back in later today. I'm so worried!
Speaker B: Come on, don't worry, you know what they say, _____

Please choose the best expression to finish the conversation:

- there's milk of human kindness by the quart in every vein!
- no need to milk it!
- no use crying over spilt milk!
- we're living in the land of milk and honey!

Q5.6. **Speaker A:** We had such a panic last week when the clients from Germany visited. My car broke down, the hotel had double booked, we had a lot of employees away sick!
Speaker B: That's terrible, so what happened?
Speaker A: Thankfully, I was able to get a taxi and sort out everything with the hotel. It was actually OK with just a few staff in the office; it meant we weren't disturbed during our meeting.
Speaker B: Sounds crazy, but you know what they say, _____

Please choose the best expression to finish the conversation:

- all good things must come to an end!
- all's well that ends well!
- it's the beginning of an end!
- the end is nigh!

Section 5: Part B

Instructions

In part B you will type in your own answers.

You should write expressions like the ones provided in the previous section (but not these exact ones, you will need to think of other expressions).

Avoid writing simple answers such as 'it is fine' or 'you will be OK', you should aim to write expressions that carry some element of wisdom (e.g., 'There's a fine line between love and hate').

Please avoid using any of the expressions that you have just seen.

Questions

Q5.7. **Speaker A:** You know, it's funny when I think about my dad.
Speaker B: Why's that?
Speaker A: We have exactly the same habits. We both like to get up early, enjoy watching history documentaries, and I suppose we're both kind of quiet and passive most of the time.
Speaker B: Well, you know what they say, _____!

Please write an appropriate expression to finish the conversation:

Q5.8. **Speaker A:** I've lived all over the world. I was born in India but grew up in Germany. I spent some time in the USA and Australia and have been in the UK for just six months.
Speaker B: So where do you consider to be home?
Speaker A: Difficult question! But I suppose, when I think of my wife, I don't mind where I live as long as it's with her.
Speaker B: That's wonderful, you know what they say, _____!

Please write an appropriate expression to finish the conversation:

Q5.9. **Speaker A:** I'm so glad we double checked the proposal for the product design before sending it to the manufacturers
Speaker B: Why, was there something wrong in the plan?
Speaker A: Very much so! In one section we had specified completely the wrong component! If that had gone unnoticed, in three months we would be spending tens of thousands of pounds on fixing the problem!
Speaker B: Good that you spotted it, you know what they say, _____!

Please write an appropriate expression to finish the conversation:

Q5.10. **Speaker A:** I get the feeling that this project is becoming complicated
Speaker B: Why's that?
Speaker A: Well, at the meetings, everybody wants to take the lead and push their ideas. I just feel that the number of people involved is having a negative effect on progress.
Speaker B: Well, you know what they say, _____!

Please write an appropriate expression to finish the conversation:

Q5.11. **Speaker A:** I think I've lost all faith in mankind!
Speaker B: That sounds a bit extreme, what happened?
Speaker A: I just can't rely on anyone or anything. My friend keeps cancelling our meeting, my assistant at work didn't do what he's supposed to, the weather forecast said sun, it's raining! You name it, you can't predict anything!
Speaker B: Well you know what they say about this life, _____!

Please write an appropriate expression to finish the conversation:

Q5.12. **Speaker A:** It's a shame that my brother and our friend Peter are not getting along well.
Speaker B: What's the problem?
Speaker A: Well, there's always been this tension between them, I just don't think they like each other very much. It's difficult because everyone has started to take sides. I like Peter very much, but if comes down to it, I have to support my brother, he's family.
Speaker B: I understand, well you know what they say, _____!

Please write an appropriate expression to finish the conversation:

Section 6: Part A

Instructions

In part A you will answer multiple-choice questions. When we explain ideas, concepts and other things to children, we often need to compare what we are explaining with something that children would understand. For example, if we want to explain the concept of love to a child, we could say: "love is a warm, fuzzy feeling that you have for a person you like". This explanation is simple, mentions words that a child would recognise ("warm", "fuzzy", "a person you like"), and would be understandable for a child. We would not say: "love is a physical state or feeling ranging from interpersonal affection to profound pleasure". This explanation is too technical, and not understandable for a child. **For each sentence, please choose the best answer to fill-the-gap.**

Questions

Q6.1. The brain works like_____.
Please choose the best answer for describing to a child what the brain **works like:**

- a calculator
- a computer
- a (computer) monitor
- a television

Q6.2. An electric current running through a wire is like _____.
Please choose the best answer for describing to a child what an electric current running through a wire **is like:**

- mice in a pipe
- a snake in a pipe
- water in a pipe
- peas in a pipe

Q6.3. A disease in the body behaves like_____.
Please choose the best answer for describing to a child what a disease in the body **behaves like:**

- an army on the attack
- a transport system
- a tourist in a city
- a shopper in a shopping mall

Q6.4. Lava running down the side of a volcano moves like_____.

Please choose the best answer for describing to a child what lava running down the side of a volcano **moves like:**

- orange juice
- blackcurrant cordial
- syrup
- jam

Q6.5. Eye lids function like _____.

Please choose the best answer for describing to a child what eye lids **function like:**

- doors
- shutters/blinds
- floors
- windows

Q6.6. Using letters to spell words is like _____

Please choose the best answer for describing to a child what using letters to spell words **is like**:

- fitting the pieces of a jigsaw puzzle together
- counting pieces of money (coins)
- moving pieces in a game of chess
- eating pieces of food

Section 6: Part B

Instructions

In part B you will type in your own answers.

Questions

Q6.7. Thunder sounds like _____

Please type in something suitable to describe to a child what thunder **sounds like:**

Q6.8. Clouds function like_____

Please type in something suitable to describe to a child what clouds **function like:**

Q6.9. The stomach functions like_____.

Please type in something suitable to describe to a child what the stomach **functions like:**

Q6.10. The ozone layer functions like_____

Please type in something suitable to describe to a child what the ozone layer **functions like:**

Q6.11. The heart functions like_____.

Please type in something suitable to describe to a child what a heart **functions like:**

Q6.12. The roots of a plant function like_____.

Please type in something suitable to describe to a child what the roots of a plant **function like:**

Section 7: Part A

Instructions

In part A you will answer multiple-choice questions. Imagine you are sitting with someone you have just met for the first time. You will be given a description of a situation and required to complete a comment. **Please choose the best option for showing the other person how you are feeling.** Note: you should avoid choosing options that make the sentence hard to understand, even if they seem correct to you. Example of a bad answer: The film was as sad as _____Mike_____. This is not a good answer, because the person you are sitting with wouldn't know who Mike is, or whether he is a sad person, a happy person, an angry person and so on.

Questions

Q7.1. Your brother is very disorganised, which you hate. Please choose the best comment to show how you feel about your brother's disorganisation:

Let me tell you about my brother, his bedroom reminds me of _____.

- a wastepaper basket
- a rubbish tip
- a dustbin
- a recycle bin

Q7.2. You thought that the party you attended last night was very boring. Please choose the best comment to show how you feel about the party:

That party was about as interesting as _____.

- watching paint crack
- watching the wall get painted
- watching paint dry
- watching paint drip

Q7.3. Last night you went to see a choir perform at a large venue. The music and harmonies were incredible. You didn't know a choir could sing this well. Please choose the best comment to show how you feel about the choir's sound:

The choir I heard last night were amazing. Their sound was like _____.

- angels rejoicing
- angels praying
- angels mourning
- angels speaking

Q7.4. You are interviewing applicants for a job. The best applicant is clearly Kate. She is outstanding and much better than the others. Please choose the best comment to show how you feel about Kate:

We've interviewed several applicants so far, but there is one lady who is clearly _____.

- the front walker
- the front of the organisation
- the front runner
- the official front

Q7.5. You think that the decision to invest in Global Enterprises LTD would lead to disaster because people in that company are very uncooperative and difficult to work with. Please choose the best comment to show how you feel about potentially working with Global Enterprises:

Working with Global Enterprises would be like _____.

- trying to get sheep to sit together
- trying to get sheep to eat
- trying to drive past a field of sheep
- trying to get sheep to make noise

Q7.6. Your little niece is always jumping around and is full of energy. You adore children, especially the fact that they are full of life. Please choose the best comment to show how you feel about your little niece:

My niece is so energetic, she's like a little _____.

- beetle
- puppy
- mouse
- bird

Section 7: Part B

Instructions

In part B you will type in your own answers. Remember, you are sitting with someone you have met for the first time and telling your comments to them. **You should** write something that shows how you feel. For example: 'The film was as scary as walking in the forest at night, alone!'. This is a good answer because the person would be able to understand how you are feeling. **You should avoid** writing things that the other person would not understand or know about. For example: 'The film was as scary as Angela'. This is not a good answer because the person you have just met would not know Angela, so would not know how scary she is!

Questions

Q7.7. You recently watched a documentary on insects in the Amazon rainforest. You were amazed by how complex and organised the life of insects is. Please complete this comment to show how amazed you were by the documentary:

I was so impressed by the complexity of life of those insects. It was like watching _____.

Q7.8. Your best friend is a very gifted athlete. She has just set a club record for the 100 metres and is focusing on training for the Olympic trials next year. You love watching her compete. Please complete this comment to show how you feel about your friend's running:

My friend is one of the best sprinters in the country. When she runs at full speed, it's like watching _____.

Q7.9. You feel that all sandwiches from Nancy's shop are the same, boring and tasteless! Please complete this comment to show how you feel about the sandwiches form Nancy's shop:

Sandwiches from Nancy's shop are about as tasty as _____.

Q7.10. The Smith project was a very successful project you did in 1992. You worked so hard on the Smith Project and are very proud of what you achieved with it. Please complete this comment to show how you feel about the Smith Project:

When I think of everything I've done over the years, I always come back to the Smith Project as my favourite. That project is really my _____.

Q7.11. You are watching a football game. Your team keep having chances to score but they miss every time. They have just missed a goal for the tenth time. Please complete this comment to show how you feel about the players in your team at the moment:

At the moment, the players are about as useful as_____.

Q7.12. Your colleague Michelle is very unkind and nasty. She spreads untrue rumours about people in the office. Please complete the comment to show how you feel about Michelle:

Michelle is about as nice as_____.

Section 8: Part A

Instructions

In part A you will answer multiple-choice questions. An idiom is a fixed phrase with a special meaning (for example: you're pulling my leg = you're joking). We often use idioms in a **fixed** way. But sometimes we **extend and play with** idioms to emphasise something or make a joke. For example: Original idiom (set phrase): He kicked the bucket (= he died). Extended idiom: He kicked the bucket so hard that it flew out of the garden! (= he died very dramatically). **For each question in this section, please choose the best option for extending the idiom.**

Questions

Q8.1. (Original idiom: it's raining cats and dogs = it's raining a lot)

Extended idiom: **It's been raining cats and dogs for so long that**_____

- the street has become flooded!
- we've been forced to call the zoo!
- we've been forced to call the stray animal collection agency!
- the street has turned into a wildlife park!

Q8.2. (Original idiom: to get a taste of your own medicine = to receive the same unpleasant experience that you yourself have given to someone else)

Extended idiom: **He got such a taste of his own medicine that** _____

- he exceeded the recommended daily dosage!
- he finally understood why everyone was upset with him!
- he finally understood medical science!
- he didn't read the label on the back!

Q8.3. (Original idiom: to be stuck between a rock and a hard place = to be in a very difficult situation)

Extended idiom: **It was a difficult decision. We were so stuck between a rock and a hard place that** _____

- we were getting very worried!
- our feet were going soft!
- our feet were beginning to resemble fossils!
- we were falling into the ground!

Q8.4. (Original idiom: break a leg! = do your best!)

Extended idiom: **Don't worry, your performance will be great! Just go out and break a leg. In fact, go out and____**

- do the very best you can!
- do something that gets you injured!
- come back with crutches!
- see where you can break your leg!

Q8.5. (Original idiom: the ball is in your court = it's your turn to respond or take action)

Extended idiom: **After her email the ball was in my court. I was expected to return it, but the problem was that** __

- I didn't want to play anymore!
- I wasn't able to make a proper booking!
- I wasn't ready to make the next decision!
- I couldn't hit the ball!

Q8.6. (Original idiom: to hit the nail on the head = to identify exactly what is causing an issue or problem)
Extended idiom: **When he said that, he hit the nail on the head so hard that**_____

- he fully explained the problem to us!
- we all felt it go through the wood!
- we saw his head start bleeding!
- he bought his own hammer!

Section 8: Part B

Instructions

In part B you will also be extending idioms, but here you will type in your own answers. You should extend the idiom so that it makes sense to someone you have just met. E.g., (Original idiom: he kicked the bucket = he died). **Your extended idiom**: he kicked the bucket so hard that _____. **A good answer** = ...it flew out of the garden! This extends the idea of a bucket being kicked. Here, the idea is that someone is kicking the bucket (i.e., dying) dramatically. Logically, a bucket kicked very hard would fly across the garden and possibly out of it. **A bad answer** = ...it froze on his foot! This would be a bad answer because it is not clear how this extends the idea of dying dramatically. Getting your foot frozen is not a logical result of kicking a bucket.

Questions

Q8.7. (Original idiom: cross that bridge when you come to it = wait to deal with a problem only if or when it happens)
Extended idiom: **Please cross that bridge when you come to it. Although, since the decision seems likely, my advice is to** _____**!**

Please extend the idiom:

Q8.8. (Original idiom: to beat around the bush = to avoid answering a question or make a clear point when talking)
Extended idiom: **He beat around the bush for so long that**_____**!**

Please extend the idiom:

Q8.9. (Original idiom: to take the cake = to be outstanding, either in a very good or a very bad way)
Extended idiom: **His comment really took the cake. In fact, it didn't just take the cake, it** _____**!**

Please extend the idiom:

Q8.10. (Original idiom: to make a mountain out of a molehill = to make a small problem seem very dramatic or important)
Extended idiom: **He made such a mountain out of a molehill that** _____**!**

Please extend the idiom:

Q8.11. (Original idiom: sitting on the fence = not making a decision about something)
Extended idiom: **He seems to be sitting on the fence about it. In fact, he's been sitting on the fence so long that_!**

Please extend the idiom:

Q8.12. (Original idiom: to fall head over heels in love = to be very much in love)
Extended idiom: **She fell so head over heels in love that** _____**!**

Please extend the idiom:

Section 9: Part A

Instructions

In part A you will answer multiple-choice questions. In conversations we often talk about one thing as if it were another thing. For example, we often talk about anxiety as a bad place or a ferocious animal. E.g., Mentally, I'm not in a good place right now (this links anxiety to a bad place). This worry is eating away at me! (this links anxiety to a ferocious animal). We often use a mixture of ideas or even the same idea throughout a **whole** conversation. This is like having a conversation in 'code'. In this section you will choose answers so that you **have whole conversations in 'code'** (e.g., about anxiety as a bad place or a ferocious animal). All the conversations are taking place online (via social media). Please click to the next page.

Questions

Scene 1. You are having a conversation online (via social media) with your friend Mary. Mary's children are with her in the room and are reading what she is typing. They can read and understand some words, but they don't understand many expressions. Mary doesn't want her children to understand the conversation, so she is writing everything in 'code'.

For each question, please choose the best response to keep Mary's 'code' (and the conversation) going.

Q9.1. Mary: Hey! It's Mary, I've got great news. The kids are reading so I'll tell you in code :)…you know I've been really hungry these past few weeks? Well today the doctor confirmed that I'm eating for two now ;)
You: Hi Mary, Wow! So you're telling me that _____.

- you've burnt your toast??!!
- you've been baking bread??!!
- you've got a bun in the oven??!!
- you've become one sandwich short of a picnic??!!

Q9.2. Mary: Yep that's right :D The stork will be paying me a visit around March 15th next year :)
You: Great! That's fantastic news! What about gender? Will you_____

- be buying pink or blue?
- be wanting yellow or orange?
- be getting it in black and white?
- be asking for green or red?

Q9.3. Mary: I don't know yet, it's far too early, but I'll be announcing it formally in a couple of weeks.
You: That's wonderful, I'm so glad to hear that once again you'll be _____

- holidaying with the family :)
- telling the family :)
- extending the family :)
- naming the family :)

Scene 2. You are having a conversation online (via social media) with your friend John. John has just had an interview for a new job. He is keeping this job a secret from almost everyone, especially the team at his current work. He is writing to you now in "code" in case someone behind him is reading what he is writing.

For each question, please choose the best response to keep John's 'code' (and the conversation) going.

Q9.4. John: Hi, it's my lunch break… On my laptop so need to write covertly in case anyone walks past and glances at the screen :)…you remember 'operation C'?
You: Hi John, haha yes I remember the famous 'operation C'! How _____

- are the gadgets working?
- has the operation been organised?
- is the operation unfolding?
- has it been to shoot a gun?

Q9.5. John: Well I've been in to assess the lay of the land, me and some rival agents met with a strict panel of drill Sergeants if you know what I mean :) It seems they've chosen their James Bond, yours truly ;)
You: Wow, that's excellent news! So you are saying_____

- you'll be given a gun soon?
- you'll be going undercover soon?
- you'll be given a car with gadgets soon?
- you'll be allied to a different government soon?

Q9.6. John: That's right. To be honest, I'm a bit worried about how to switch over from my current operation if you catch my drift :) The crew and captain will not be very pleased that I'm jumping ship!
You: Don't worry; it'll be fine. Think of it like this: _____

- every gadget is useful!
- every operation comes to an end!
- every agent loses a few gadgets!
- every operation costs money!

Section 9: Part B

Instructions

In part B you will also be continuing "coded" conversations. Here, you will need to type in your own answers. Here is an example: **Example scene.** You are having a conversation online (via social media) with your friend Anna who is at work. Anna has told you in secret that she likes one of her colleagues romantically, and she thinks that he likes her too. She doesn't want anyone to know about this, so she is writing to you in "code" in case someone walks by and looks at her computer screen. **Anna:** You remember the Shakespearean story I told you about?
You: Oh yes! Let me ask, _____. **A good response =** how is the story of Romeo and Juliet going? :) This is a good response because: 1) your friend would understand what you are talking about and 2) you have kept the 'code' going. **A bad response** = do you know if your colleague is in love with you too? This is a bad response because: 1) it is too direct and 2) it doesn't keep the "code" going. Please click to the next page.

Questions

Scene 3. You are having a conversation online (via social media) with your mum. You are talking about your brother Jack. Jack is an energetic guy who always gets up early and never seems to run out of energy. You and your mum are writing to each other in "code" to make a joke about Jack.

Please write responses in 'code' that keep the conversation with your mum going.

Q9.7. Your mum: Jack, the machine called in earlier!
You: Haha, I know we joke about it, but it's really true; he is a machine! You can always see that he is_____

Q9.8. Your mum: You'll never believe it, he steamed over to the house in search of midday fuel, again!
You: That sounds about right! Even though he left home several years ago, he still comes here for refuel. Why didn't you just tell him_____?

Q9.9. Your mum: Well, it was quite nice to feel like the mechanic again, or at least the petrol station attendant! He actually seemed a bit conked out.
You: Really, well, I'm sure that after receiving his refuelling and a bit of home mechanics, he's now _____!

Scene 4. You are having a conversation online (via social media) with your colleague Peter. The two of you are talking about another colleague of yours who is great at his job and is due to receive a surprise award. You both enjoy speaking in "code" about this employee.

Please write responses in 'code' that keep the conversation with Peter going.

Q9.10. Peter: Have you heard? The wizard has done his magic again. I'm talking about the secret magic award.

You: Oh yes, I heard Mr magic is due to be _____

Q9.11. Peter: Yes, that's right, his spells have been creating quite a positive stir in the kingdom

You: Which spell in particular? Will the magic circle commend him for_____?

Q9.12. Peter: I think his main magical achievement was something like that. But he's really all-round enchanting; he's simply been running our show for a long time

You: I agree, I'm completely _____

Appendix C Consent form for Chinese participants

THE UNIVERSITY of York

DEPARTMENT OF EDUCATION

Name of Researcher: David O'Reilly, PhD candidate
Study: Metaphoric competence and vocabulary knowledge study

Brief Description of Study

The aim of this study is to explore a concept called 'metaphoric competence' in relation to vocabulary knowledge. I would like to ask you to do four sets of questions (tests). You can either do everything on your computer at home in your own time or you can arrange to come and do the tests with me present. The tests include: a) a metaphoric competence test that I have designed; b) a breadth of vocabulary knowledge test; c) a depth of vocabulary knowledge test, d) the Oxford Online Placement test. Test a) should take 1 hour 30 minutes, test b) around 10 minutes and test c) around 20 minutes and test d) around 30 minutes. In total, this should be about 2 hours 30 minutes (although it is fine to take longer if you need).

In exchange for taking part, you will receive a £5 reward (£4 Amazon voucher plus £1 cash). You will also be invited to a feedback session in which we go through the correct answers and discuss the tests. If you are not able to attend this, I will arrange feedback by Skype, email or phone.

The information I obtain from you will help me in my PhD research. Some of the data I collect from you will be presented in my PhD thesis and (potentially) at conferences; however, your identity will be coded and kept anonymous. (Only I will have access to identifiable data). You are free to stop your participation at any point of the study. There will be no negative consequences for you should you do so. If you wish to remove your data, please let me know by 1st October 2015, as after this your data will anonymised and incorporated into reports and so difficult to remove.

If you have any further questions about the study, or would like a debrief after the study is completed, please write to david.oreilly@york.ac.uk. For any concerns of complaints please contact the researcher's supervisor and Chair of the Education Ethics Committee at emma.marsden@york.ac.uk or the PhD in Education programme leader at chris.kyriacou@york.ac.uk.

INFORMED CONSENT (Metaphoric competence and vocabulary knowledge study)

I have read the statement concerning the research that I am being asked to take part in, and I have had the opportunity to ask questions. I understand that I may withdraw at any time, and that my identity will be kept anonymised if the PhD is published or presented at conferences. I am happy to take part in the research.

Signed …… Date ……..
Name …… Native language……

Appendix D Rating scale outliers

Rating scale item outliers: Test 3-Vehicle Acceptability-R

			NS group (*n* = 31)			
Item no.	Item content	Res-designed accept.	*M* (%)	*SD* (%)	Item kept	Rating needed to score '1' (*correct*)
1	His blood began to boil as he started shouting	high	98.7	4.8	yes	94-100%
2	He slipped into a depression	high	87.7	25.7	no	
3	His body went fat after a few years	low	32.2	30.8	no	
4	The whole theory fell apart	high	98.7	5.5	yes	93-100%
5	The project is going ahead as planned	high	99.7	1.4	yes	98-100%
6	He couldn't bottle his anger up anymore so he started shouting	high	96.5	9.7	yes	87-100%
7	It was an attractive proposal	high	96.6	13.2	yes	83-100%
8	The idea holds up in principle	high	92.8	18.4	yes	74-100%
9	To her the drunken man was repulsive	high	95.9	11.1	yes	85-100%
10	The theory was the colour of brick	low	7.2	13.5	yes	0-21%
11	There was a lot of electricity between the dog and ball	low	19.9	26.0	no	
12	Her hair had almost arrived at being grey	low	8.2	13.1	yes	0-21%
13	We entered the front door of the plan	low	9.6	20.0	yes	0-30%
14	He bubbled as he began shouting	low	22.0	30.8	no	
15	Their similarities jerked them together	low	12.4	21.8	yes	0-34%
16	She turned orange as she started shouting at him	low	9.4	22.4	yes	0-32%
17	He freshened his ideas	low	39.7	37.7	no	
18	He told a white lie	high	100.0	0.0	yes	100%
19	She made a firm proposal to the client	high	95.0	18.6	yes	76-100%
20	He tried to pull the wool over my eyes	high	99.9	0.5	yes	99-100%
21	He never has time to shoot the breeze	high	69.8	41.2	no	
22	He has a killer headache	high	79.4	32.2	no	
23	We solved the teased out problem very easily	low	18.2	28.0	no	
24	I picked up a job last week	high	74.0	33.8	no	
25	The comment blunts	low	2.8	5.2	yes	0-8%
26	We asked for a called day at 6pm	low	7.5	18.1	yes	0-26%
27	I will give you a show of the ropes tomorrow	low	21.9	29.7	no	
28	The team are trained to makes calls coldly; customers never expect their calls!	low	14.6	23.4	yes	0-38%

Rating scale item outliers: Test 4-Topic/Vehicle-R

Item no.	Item content	Options	Res-designed rank	Rank	*M* (%)	*SD* (%)	Item kept	Rating of rank 1 option needed to score '1' (*correct*)
		NS group 1 (*N* = 15)						
1	The CCTV cameras are the ____ of the building.	a) eyeballs		2	48.3	33.4	yes	
		b) goggles		4	8.3	8.6		
		c) glasses		3	14.3	14.3		
		d) eyes	1	1	95.9	10.4		86-100%
2	New products at the end of a long production process are the ____ of large companies.	a) acorns		4	9.2	16.6	yes	
		b) vegetables		2	13.4	22.6		
		c) fruits	1	1	96.5	7.1		89-100%
		d) seeds		3	9.9	20.4		
3	This park is the ____ of our city.	a) lungs	1	1	82.5	34.5	yes	48-100%
		b) kidneys		4	3.8	7.0		
		c) mouth		3	11.3	17.5		
		d) chest		2	12.7	19.5		
4	The main argument is the ____ of the essay.	a) bread		2	14.9	19.5	yes	91-100%
		b) meat	1	1	97.3	5.9		
		c) pasta		4	2.5	3.8		
		d) rice		3	3.8	4.8		
5	The company's internal mail team are the ____ of the organisation.	a) brain	1	2	62.2	39.2	No	_
		b) fingers		3	13.1	20.0		
		c) blood		1	58.5	41.7		
		d) skin		4	4.9	7.8		
6	The bee hive is the ____ of the animal kingdom.	a) airport	1	1	47.3	37.9	yes	9-85%
		b) bus station		3	24.2	29.7		
		c) train station		2	26.2	32.9		
		d) taxi rank		4	18.1	30.2		
		NS group 2 (*N* = 16)						
7	Volcanoes are the ____ of the earth.	a) mouths		3	40.4	37.9	yes	
		b) bruises		4	15.6	20.7		
		c) blisters		2	46.0	32.2		
		d) pimples	1	1	48.8	39.5		9-88%
8	Chemical elements are the ____ of life.	a) stones		3	22.3	28.9	yes	
		b) chains		2	38.4	33.1		
		c) building blocks	1	1	96.1	5.9		90-100%
		d) roof tiles		4	6.3	9.8		
9	The sales team are the ____ of the organisation.	a) shepherds		2	30.2	29.0	yes	45-100%
		b) bakers		4	15.9	24.5		
		c) farmers		3	19.2	21.2		
		d) hunters	1	1	77.9	32.8		
10	Killer whales are the ____ of the sea.	a) hyenas		3	23.3	31.4	yes	25-100%
		b) horses		4	18.9	29.8		
		c) rhinos		2	31.9	34.5		
		d) wolves	1	1	62.5	37.9		
11	The outside walls are the ____ of the building.	a) lips		4	5.25	7.1	yes	66-100%
		b) skin	1	1	86.1	20.4		
		c) ears		2	20.0	21.5		
		d) head		3	14.6	21.7		
12	Alcohol is the ____ of the drunk person.	a) steering wheel		2	40.0	27.6	yes	59-100%
		b) fuel	1	1	83.6	24.7		
		c) engine		3	32.9	29.7		
		d) trunk/bonnet		4	3.2	3.3		

Appendix E Participant outliers

Participant outliers in the raw scores data

Ppt.	Data file outlier	Group	Test	Score	Out of	Problem
N3A	NS	1	MC Test 1-Phrasal Verbs-R	9	10	< 9.16
N10B	NS	1+2	MC Test 1-Phrasal Verbs-P	7	10	< 7.16
18A	NNS	1&2	MC Test 2-Metaphor Layering (Ab)-R	0	6	< 0.54
18A	NNS	1&2	MC Test 2-Metaphor Layering (Ab+B+C)-R	2	18	< 2.21
18A	NNS	1&2	MC Test 2-Metaphor Layering (Aa+Ab+B+C)-R	3	24	< 3.15
N2B	NS	1&2	MC Test 2-Metaphor Layering (Aa)-R	0	6	< 0.87
N2B	NS	1&2	MC Test 2-Metaphor Layering (Aa+Ab)-R	4	12	< 5.64
N15B	NS	1&2	MC Test 2-Metaphor Layering (B)-R	3	6	< 3.02
N15B	NS	1&2	MC Test 2-Metaphor Layering (B+C)-R	6	12	< 6.12
18A	NNS+NS	1&2	MC Test 2-Metaphor Layering (Ab)-R	0	6	< 0.84
N2A	NS	1&2	MC Test 3-Vehicle exploitation-R	5	12	< 5.99
N2A	NS	1&2	MC Test 3-Vehicle word class-R	3	6	< 3.53
N2A	NS	1&2	MC Test 3-Vehicle Acceptability-R	8	18	< 10.09
4A	NNS	1+2	MC Test 4-Topic/Vehicle-R	0	6	< 0.02
4A	NNS+NS	1	MC Test 4-Topic/Vehicle-R	0	6	< 0.14
4A	NNS+NS	1+2	MC Test 4-Topic/Vehicle-R	0	6	< 0.14
17B	NNS	1+2	MC Test 5-Topic Transition-R	1	6	< 1.22
N2A	NS	1	MC Test 5-Topic Transition-R	5	6	< 5.16
31A[a]	NNS+NS	1	MC Test 5-Topic Transition-R	2	6	< 2.04
17B	NNS+NS	1+2	MC Test 5-Topic Transition-R	1	6	< 1.46
N15A	NS	1+2	MC Test 5-Topic Transition-P	4	12	< 4.30
N7B	NS	1+2	MC Test 6-Heuristic-R	3	6	< 3.29
N16B	NS	1+2	MC Test 7-Feelings-R	3	6	< 3.57
N13A	NS	1+2	MC Test 8-Idiom Extension-P	2	12	< 2.19
N15B	NS	1+2	MC Test 9-Metaphor Continuation-P	0	12	< 1.05
23A	NNS	1&2	VYesNo	2100	10000	< 2176
N2B	NS	1&2	VYesNo	3074	10000	< 3890
13B	NNS	1&2	WAT	89	160	< 93
N10B	NS	1&2	WAT	115	160	< 117
36B	NNS	1&2	OOPT Listening	119 (C2+)	120 (C2+)	> 119
46A	NNS	1&2	IELTS Writing	7.5	9	> 7.47
50A	NNS	1&2	IELTS Writing	7.5	9	> 7.47
56A	NNS	1&2	IELTS (overall)	5.0	9	< 5.10
12B	NNS	1&2	IELTS (overall)	5.0	9	< 5.10

[a] Removed from the NNS+NS data file but not the NNS data file. Because 31A was not an outlier in the NNS data file, they were also not removed for the distractor analysis (section 5.2.5).

GUIDE TO COLUMNS

Participant code refers to the identity given to participants by the researcher.

Data file outlier refers to the data file in which the participant is an outlier. For instance, 18A is a participant outlier for MC Test 2-Metaphor Layering (Ab)-R in both the NNS data file (i.e., in relation to NNS scores only), and the NNS+NS data file (i.e., in relation to NNS and NS scores considered together). Participants who were outliers in the NNS or NS data files were also removed from the NNS+NS data file. However, participants who were outliers in the NNS+NS data file only were not removed from the NNS or NS data files provided they were not outliers in these files.

Group shows which group's data the participant is an outlier with respect to. Values '1' and '2' indicate that the participant is an outlier in relation to group 1 or group 2's data only; '1&2' indicates that groups 1 and 2 encountered exactly the same items; '1+2' indicates that the participant is an outlier in relation to a group 1 and 2's scores combined, with both groups having encountered different items.

Test indicates which test the participant is an outlier for. Receptive and productive tests are treated separately and are tagged –R and –P. For MC tests 2 and 3 outliers were identified in relation to both the overall test and component parts (e.g., Ab questions only in Test 2).

Score indicates the raw score the participant obtained for the test (or section).

Out of shows the maximum possible score for that test (or section).

Problem lists the reason why a participant is an outlier. In most cases, this is due to their score falling below the group mean minus three standard deviations.

Appendix F Final sets of items retained in the NNS, NS and NNS+NS data files

MC Items Retained after Data Cleaning Process (Chapter 5)

			NNS data file		NS data file		NNS+NS data file	
MC test	R/P	Group	*K*	Items retained	*K*	Items retained	*K*	Items retained
T1-Phrasal Verbs	R	1	4	1,4,5,7	10[b]	1,2,3,4,5,6,7,8,9,10	9[a]	1,3,4,5,6,7,8,9,10
		2	7	11,12,13,14,15,16,20	10[b]	11,12,13,14,15,16,17,18,19,20	8	12,13,14,15,16,17,18,20
	P	2	4	1,2,4,7	10[b]	1,2,3,4,5,6,7,8,9,10	8	2,3,4,5,6,7,8,10
		1	4[a]	13,17,18,20	10[b]	11,12,13,14,15,16,17,18,19,20	7[a]	12,15,16,17,18,19,20
T2-Metaphor Layering	R	1&2	11	1a,1b,3a,4a,4b,5a,5b, 6a,6b,B7,C15	6	1a,4a,5a,B7,C15,C17	16	1a,1b,3a,4a,4b,5a,5b,6b, B7,B8,B10,B11,B12,C14,C15,C17
T3-Vehicle Acceptability	R	1&2	10	1,4,5,6,7,8,9,18,19,20	11	1,4,6,7,8,10,12,13,15,19,25	18	1,4,5,6,7,8,9,10,12,13,15, 16,18,19,20,25,26,28
T4-Topic/Vehicle	R	1	4	1,2,3,4	4	1,2,3,4	4	1,2,3,4
		2	4	8,9,11,12	4	8,9,11,12	4	8,9,11,12
	P	2	4	1,2,3,5	4	1,2,3,5	4	1,2,3,5
		1	4	7,8,9,11	4	7,8,9,11	4	7,8,9,11
T5-Topic Transition	R	1	4	3,4,5,6	6[b]	1,2,3,4,5,6	4	3,4,5,6
		2	4	7,10,11,12	6[b]	7,8,9,10,11,12	4	7,10,11,12
	P	2	6	1,2,3,4,5,6	4	2,3,4,5	6	1,2,3,4,5,6
		1	4	7,8,11,12	4	8,9,10,11	6	7,8,9,10,11,12
T6-Heuristic	R	1	5	2,3,4,5,6	6[c]	1,2,3,4,5,6	6	1,2,3,4,5,6
		2	5	7,8,9,10,11	6[b]	7,8,9,10,11,12	4	7,8,10,11
	P	2	4	1,4,5,6	4	1,4,5,6	4	1,4,5,6

		1	4	7,8,9,11	4	8,10,11,12	5	7,8,9,11,12
T7-Feelings	R	1	5[a]	2,3,4,5,6	6[b]	1,2,3,4,5,6	5[a]	1,3,4,5,6
		2	6[a]	7,8,9,10,11	6[b]	7,8,9,10,11,12	6[a]	7,8,9,10,11,12
	P	2	4	2,4,5,6	4	1,2,3,5	5	1,2,4,5,6
		1	4	8,9,11,12	4	7,8,9,10	4	8,10,11,12
T8-Idiom Extension	R	1	4	1,2,3,4	5	2,3,4,5,6	5	2,3,4,5,6
		2	4	9,10,11,12	4	7,10,11,12	5	8,9,10,11,12
	P	2	6	1,2,3,4,5,6	6	1,2,3,4,5,6	6	1,2,3,4,5,6
		1	6	7,8,9,10,11,12	4	9,10,11,12	6	7,8,9,10,11,12
T9-Metaphor Continuation	R	1	4	1,2,3,5	5[b]	1,2,3,4,6	5	1,2,3,4,6
		2	4	8,10,11,12	5	7,8,9,10,11	6[a]	7,8,9,10,11,12
	P	2	4[a]	1,4,5,6	4	1,2,3,6	4[a]	1,3,5,6
		1	5	8,9,10,11,12	5	7,8,10,11,12	5[a]	7,9,10,11,12
MC Test Battery[e]	R	1	50	items above	59	items above	72	items above
		2	54	items above	58	items above	71	items above
	P	1	33	items above	35	items above	37	items above
		2	32	items above	36	items above	37	items above
	R & P	1	83	items above	94	items above	109	items above
		2	86	items above	94	items above	108	items above

Note. Key to column headings: R/P = receptive or productive, K = number of test items retained at end of analysis.

[a] Items chosen to ensure no differences between G1 & G2 scores.

[b] Items for which all participants scored full marks retained.

[c] Reasons [a] and [b].

[d] Before Test 4-Topic/Vehicle-P was deleted from the MC Test Battery, items 7,8,9,11 (Group 1) and 1,2,3,4 (Group 2) had been retained.

[e] Reliability estimates of all items retained above from tests 1-9.

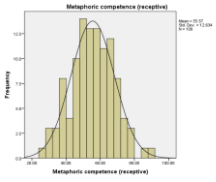
Appendix G EFA of NNS data: Data screening

NNS Data File: Normality, Skewness and Kurtosis of Variables

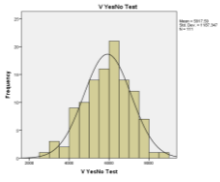
						Kolmogorov-Smirnov			Shapiro-Wilk		
Test/variable	*N*	Skw.	*SE* of Skw.	Kurt.	*SE* of Kurt.	Statistic	*df*	Sig.	Statistic	*df*	Sig.
T1-Phrasal Verbs-R	112	-0.29	0.23	-0.75	0.45	0.16	112	.00	0.94	112	.00
T1-Phrasal Verbs-P	112	-0.30	0.23	-0.40	0.45	0.19	112	.00	0.90	112	.00
T2-Metaphor Layering-R	111	0.02	0.23	-1.04	0.46	0.08	111	.05	0.96	111	.00
T3-Vehicle Acceptability-R	112	-0.24	0.23	-0.75	0.45	0.09	112	.04	0.97	112	.01
T4-Topic/Vehicle-R	111	-0.32	0.23	-0.45	0.46	0.20	111	.00	0.91	111	.00
T5-Topic Transition-R	111	-0.60	0.23	-0.21	0.46	0.20	111	.00	0.87	111	.00
T5-Topic Transition-P	112	0.43	0.23	-0.63	0.45	0.14	112	.00	0.95	112	.00
T6-Heuristic-R	112	-0.43	0.23	-0.68	0.45	0.21	112	.00	0.91	112	.00
T6-Heuristic-P	112	-0.36	0.23	-0.55	0.45	0.13	112	.00	0.95	112	.00
T7-Feelings-R	112	-0.15	0.23	-0.33	0.45	0.10	112	.01	0.97	112	.01
T7-Feelings-P	112	0.30	0.23	-0.66	0.45	0.15	112	.00	0.96	112	.00
T8-Idiom Extension-R	112	1.10	0.23	0.16	0.45	0.27	112	.00	0.78	112	.00
T8-Idiom Extension-P	112	0.59	0.23	-1.05	0.45	0.19	112	.00	0.86	112	.00
T9-Metaphor Continuation-R	112	-0.01	0.23	-0.90	0.45	0.16	112	.00	0.92	112	.00
T9-Metaphor Continuation-P	112	-0.02	0.23	-1.16	0.45	0.13	112	.00	0.94	112	.00
VYesNo	111	-0.27	0.23	-0.23	0.46	0.05	111	.20	0.99	111	.56
Word Associates Test	111	-0.32	0.23	-0.17	0.46	0.07	111	.20	0.98	111	.20
OOPT Use of English	112	-0.52	0.23	0.07	0.45	0.08	112	.12	0.98	112	.04
OOPT Listening	112	-0.07	0.23	-0.41	0.45	0.09	112	.04	0.98	112	.04
IELTS Reading	111	0.00	0.23	-0.32	0.46	0.16	111	.00	0.95	111	.00
IELTS Writing	111	0.73	0.23	1.73	0.46	0.32	111	.00	0.84	111	.00
IELTS Speaking	111	0.23	0.23	-0.29	0.46	0.22	111	.00	0.92	111	.00
IELTS Listening	111	-0.07	0.23	-0.64	0.46	0.14	111	.00	0.96	111	.00

Histograms: Most Extreme Examples of Normality, Nonnormality, Skewness and Kurtosis

Normal

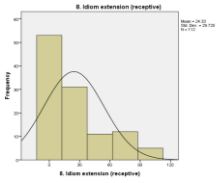


MC-R

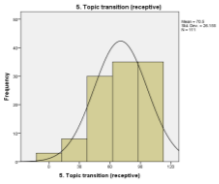


VYesNo

Nonnormal

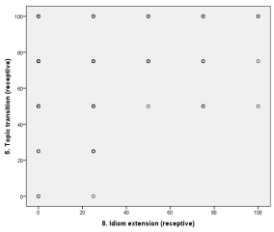


Test 8-Idiom Extension-R

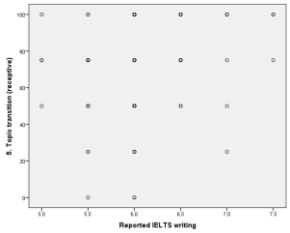


Test 5-Topic Transition-R

Scatterplots: Most likely Variables with Nonlinearity and Heteroscedasticity



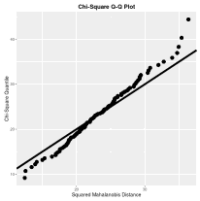
Test 8-R (x axis) and Test 5-R (y axis)



IELTS Writing (x axis) and Test 6-R (y axis)

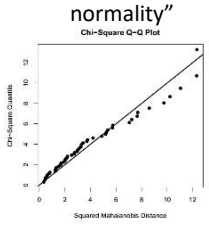
Multivariate normality: Comparison of Present Study Chi-Square Q-Q Plot with Korkmaz, Goksuluk and Zararsiz's (2014) Example

Chi-square Q-Q plot for the NNS data



Chi-Square Q-Q Plot, Variables for NNS EFA

Korkmaz, Goksuluk and Zararsiz's (2014, p. 156) example of "possible departure from [multivariate] normality"



Example of Chi-Square Q-Q Plot from

Exploring Transformations of NNS Variables

	Initial variable					Transformed variable					
	Skewness		Kurtosis		K-S[a]		Skewness		Kurtosis		K-S[a]
Variable	Stat.	SE	Stat.	SE	Sig.	Type	Stat.	SE	StaT.	SE	Sig.
T1-Phrasal Verbs-R	-0.29	0.23	-0.75	0.45	.00	_	_	_	_	_	
T1-Phrasal Verbs-P	-0.30	0.23	-0.40	0.45	.00	_	_	_	_	_	
T2-Metaphor Layering-R	0.02	0.23	-1.04	0.46	.05	SQRT	-0.40	0.23	-0.53	0.46	.05
T3-Vehicle Acceptability-R	-0.24	0.23	-0.75	0.45	.04	Rfl & SQRT	-0.48	0.23	0.06	0.45	.09
T4-Topic/Vehicle-R	-0.32	0.23	-0.45	0.46	.00	_	_	_	_	_	
T5-Topic Transition-R	-0.60	0.23	-0.21	0.46	.00	_	_	_	_	_	
T5-Topic Transition-P	0.43	0.23	-0.63	0.45	.00	SQRT	-0.66	0.23	0.05	0.45	.00
T6-Heuristic-R	-0.43	0.23	-0.68	0.45	.00	_	_	_	_	_	
T6-Heuristic-P	-0.36	0.23	-0.55	0.45	.00	Rfl & SQRT	-0.43	0.23	-0.09	0.45	.00
T7-Feelings-R	-0.15	0.23	-0.33	0.45	.01	_	_	_	_	_	
T7-Feelings-P	0.30	0.23	-0.66	0.45	.00	_	_	_	_	_	
T8-Idiom Extension-R	1.10	0.23	0.16	0.45	.00	_	_	_	_	_	
T8-Idiom Extension-P	0.59	0.23	-1.05	0.45	.00	LOG10	-0.72	0.28	-0.68	0.55	.00
T9-Metaphor Continuation-R	-0.01	0.23	-0.90	0.45	.00	_	_	_	_	_	
T9-Metaphor Continuation-P	-0.02	0.23	-1.16	0.45	.00	Rfl & SQRT	-0.40	0.23	-0.73	0.45	.00

[a]Kolmogorov-Smirnov test, Lilliefors Significance Correction.

Appendix H EFA of NNS data, supplementary tables and figures

Eigenvalues and Mean Bootstrap Results across 5,000 Resamples

Eig. No.	*N* (boots.)	Sample eig.	*M* boot.	*SD* of boot.	*Mdn* boot.	95% CI[a] Lower	Upper
1	5000	6.39	6.47	0.51	6.48	5.61	7.30
2	5000	1.58	1.94	0.15	1.92	1.72	2.19
3	5000	1.46	1.67	0.10	1.67	1.51	1.85
4	5000	1.29	1.48	0.09	1.47	1.34	1.62
5	5000	1.22	1.32	0.07	1.32	1.20	1.45
6	5000	1.12	1.18	0.07	1.18	1.08	1.30
7	5000	1.02	1.07	0.06	1.07	0.97	1.17
8	5000	0.93	0.97	0.05	0.97	0.88	1.06
9	5000	0.86	0.88	0.05	0.88	0.80	0.96
10	5000	0.85	0.80	0.05	0.80	0.72	0.87
11	5000	0.75	0.72	0.04	0.72	0.65	0.80
12	5000	0.71	0.65	0.04	0.65	0.59	0.72
13	5000	0.66	0.59	0.04	0.59	0.53	0.66

*The 5th and 95th percentiles of bootstrapped eigenvalues.

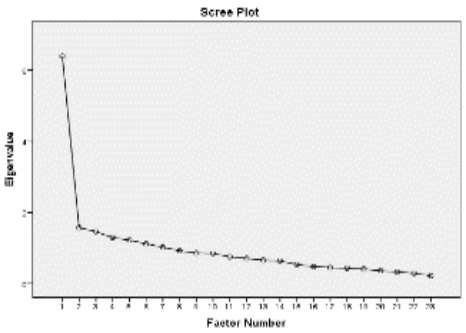
KMO Test for Individual Variables

Individual variables	MSA
T1-Phrasal Verbs-R	0.76
T1-Phrasal Verbs-P	0.79
T2-Metaphor Layering-R	0.89
T3-Vehicle Acceptability-R	0.86
T4-Topic/Vehicle-R	0.66
T5-Topic Transition-R	0.78
T5-Topic Transition-P	0.84
T6-Heuristic-R	0.82
T6-Heuristic-P	0.79
T7-Feelings-R	0.88
T7-Feelings-P	0.80
T8-Idiom Extension-R	0.79
T8-Idiom Extension-P	0.87
T9-Metaphor Continuation-R	0.72
T9-Metaphor Continuation-P	0.86
V YesNo	0.82
Word Associates Test	0.85
OOPT Use of English	0.88
OOPT Listening	0.87
IELTS Reading	0.89
IELTS Writing	0.74
IELTS Speaking	0.85
IELTS Listening	0.85

Overall KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy. (MSA)		0.835
Bartlett's Test of Sphericity	Approx. Chi-Square	753.363
	df	253
	Sig.	0.000

NNS Scree Plot



Total Variance Explained (in SPSS)

F	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	6.392	27.792	27.792	5.871	25.524	25.524	3.583
2	1.575	6.849	34.641	1.052	4.575	30.099	2.883
3	1.461	6.354	40.995	.831	3.611	33.710	2.612
4	1.293	5.620	46.615	.762	3.315	37.025	2.951
5	1.223	5.319	51.935	.624	2.715	39.740	2.629
6	1.122	4.877	56.812	.556	2.416	42.156	1.361

Note. Extraction Method: Principal Axis Factoring.
[a] When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

Total Variance Explained (in R)

	F1	F2	F3	F4	F5	F6
SS loadings	1.93	1.92	1.51	1.35	1.44	1.55
% Variance	8.39	8.35	6.57	5.87	6.26	6.74
Cumulative %	8.39	16.74	23.30	29.17	35.43	42.17

NNS Factor Correlation Matrix

	2	3	4	5	6
1. (F1) EVS	.48	.25	.26	.37	.33
2. (F2) EGC	_	.25	.2	.32	.27
3. (F3) EGMC		_	.28	.21	.31
4. (F4) EIMP			_	.29	.27
5. (F5) ETVA				_	.34
6. (F6) EMLP					_

1. (F1) EVS = English Vocabulary Size
2. (F2) EGC = English General Comprehension
3. (F3) EGMC = English Grammatical Metaphoric Competence
4. (F4) EIMP = English Illocutionary Metaphor Production
5. (F5) ETVA = English Topic/Vehicle Acceptability
6. (F6) EMLP = English Metaphor Language Play

NNS EFA Structure Matrix

Tests/Variables	F 1	F 2	F 3	F 4	F 5	F 6
IELTS Listening	0.46	0.72	0.36	0.33		
IELTS Reading	0.38	0.68			0.30	0.34
Test 2-Metaphor Layering-R	0.45	0.54		0.35	0.32	
IELTS Speaking	0.45	0.48				0.36
Test 6-Heuristic-R		0.37				
V YesNo	0.90	0.47			0.34	0.34
IELTS Writing	0.51	0.45			0.33	
Test 9-Metaphor Continuation-R	0.34					
Test 1-Phrasal Verbs-P			0.63			0.31
OOPT Listening	0.41	0.42	0.60		0.41	
Test 1-Phrasal Verbs-R		0.30	0.40			
OOPT Use of English	0.44	0.46	0.52		0.40	0.36
Word Associates Test	0.51	0.31	0.50	0.41	0.49	0.51
Test 8-Idiom Extension-P	0.33	0.31	0.31		0.34	0.81
Test 9-Metaphor Continuation-P	0.49		0.35	0.40	0.42	0.57
Test 8-Idiom Extension-R	0.35					0.42
Test 7-Feelings-R	0.39		0.37	0.38		0.45
Test 4-Topic/Vehicle-R					0.54	
Test 5-Topic Transition-R					0.57	0.32
Test 3-Vehicle Acceptability-R					0.52	
Test 6-Heuristic-P				0.82		
Test 7-Feelings-P	0.37	0.31		0.42		
Test 5-Topic Transition-P		0.37	0.33	0.42	0.36	0.38

Correlations: NNS Tests

T	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	.23**	.10	.05	.02	.04	.11	.06	-.02	.08	.08	-.06	.13	.11	.10	.17*	.21*	.24**	.27**	.27**	.02	.16*	.23**
2	_	.03	.06	.04	.05	.23**	.17*	.12	.29**	.12	.12	.24**	.09	.21*	.13	.38**	.33**	.30**	.09	-.06	.13	.18*
3		_	.18*	.12	.18*	.29**	.26**	.32**	.15	.22**	.18*	.23**	.23**	.28**	.47**	.28**	.33**	.27**	.41**	.27**	.24**	.43**
4			_	.27**	.33**	.22**	.22**	.18*	.12	.11	.14	.23**	.09	.26**	.19*	.26**	.20*	.24**	.23**	.20*	.12	.25**
5				_	.30**	.21*	.04	.17*	.14	.14	.14	.13	.19*	.16*	.11	.26**	.09	.17*	.08	.09	.09	.00
6					_	.17*	.09	.16*	.15	.11	.19*	.25**	.01	.26**	.22*	.38**	.33**	.27**	.22*	.20*	.20*	.19*
7						_	.25**	.33**	.24**	.28**	.13	.40**	.14	.38**	.16*	.33**	.36**	.28**	.27**	.15	.17*	.35**
8							_	.22**	.10	.23**	.10	.19*	.18*	.25**	.16*	.28**	.33**	.22**	.24**	.10	.21*	.29**
9								_	.31**	.36**	.14	.22*	.04	.34**	.27**	.38**	.26**	.19*	.17*	-.02	.06	.28**
10									_	.30**	.33**	.35**	.05	.37**	.36**	.40**	.26**	.33**	.27**	.10	.26**	.36**
11										_	.13	.13	.18*	.21*	.32**	.25**	.25**	.37**	.19*	.28**	.15	.31**
12											_	.40**	.09	.30**	.33**	.21*	.25**	.28**	.22*	.15	.15	.11
13												_	.05	.53**	.33**	.45**	.35**	.28**	.34**	.19*	.31**	.28**
14													_	.29**	.28**	.16	.27**	.32**	.11	.16*	.18*	.21*
15														_	.49**	.52**	.38**	.30**	.24**	.19*	.34**	.29**
16															_	.55**	.45**	.39**	.38**	.44**	.38**	.47**
17																_	.42**	.40**	.27**	.23**	.28**	.38**
18																	_	.49**	.35**	.22*	.33**	.35**
19																		_	.36**	.12	.17*	.45**
20																			_	.34**	.35**	.50**
21																				_	.39**	.26**
22																					_	.40**
23																						_

Note. Key: (1) Test 1-Phrasal Verbs-R; (2) Test 1-Phrasal Verbs-P; (3) Test 2-Metaphor Layering-R; (4) Test 3-Vehicle Acceptability-R; (5) Test 4-Topic/Vehicle-R; (6) Test 5-Topic Transition-R; (7) Test 5-Topic Transition-P; (8) Test 6-Heuristic-R; (9) Test 6-Heuristic-P; (10) Test 7-Feelings-R; (11) Test 7-Feelings-P; (12) Test 8-Idiom Extension-R; (13) Test 8-Idiom Extension-P; (14) Test 9-Metaphor Continuation-R; (15) Test 9-Metaphor Continuation-P; (16) V YesNo Test; (17) Word Associates Test; (18) OOPT Use of English; (19) OOPT Listening; (20) IELTS Reading; (21) IELTS Writing; (22) IELTS Speaking; (23) IELTS Listening.

[a] Determinant = .001.

*significant at the .05 level.

**significant at the .01 level.

Adequacy of the Six-Factor Structure: NNS EFA

Test/statistic	Criterion (source)	Present study value
Degrees of freedom (*df*) for null model	_	253
Objective function for null model	_	7.53
Chi-square for null model	_	753.36
df for model	_	130
Objective function for model	_	1.12
Root mean square of residuals (RMSR)	< .05 indicates close fit (Hu & Bentler, 1999)	0.04
df corrected RMSR	_	0.05
Harmonic number of observations	_	111
Empirical chi-square	_	82.51 w/ prob < 1
Total number of observations	_	112
Likelihood chi-square	_	110.64 w/ prob < .89
Tucker Lewis index of factor reliability	Minimum 0.95[a] (Hu & Bentler, 1999)	1.08
Root mean square error of approximation (RMSEA) index	< .05 indicates close fit (Hu & Bentler, 1999)	0
90% confidence intervals	_	NA and 0.023
Comparative fit index (CFI)[b]	> .95 indicates good fit (Hu & Bentler, 1999; Tabachnick & Fidell, 2013)	1[c]
BIC	_	-502.67
Fit based on diagonal off values	> .95 indicates good fit (A. Field et al., 2012)	0.98

[a] Since index is not normalised, values exceeding '1' are permitted (Hu & Bentler, 1999).
[b] Calculated as 1 - ((chi-square of model -*df* of model)/(chi-square of null - *df* of null)) and adjusted to 1 or 0 if above or below these ranges.
[c] Adjusted down to '1' (Kenny, 2015).

NNS EFA Reliability of Factors

Factors		Reliability (internal consistency) estimates												
		Tests-within-factor				Items-within-factor								
						Group 1			Group 2			Total[a]		
No.	Name	*K*	Tests	*r*	α	*N*	*K*	α	*N*	*K*	α	*N*	*K*	α
1	EVS	2	VYesNo IELTS Writing	.47[b]	_	55	_	_	55	_	_	110	_	_
2	EGC	5	IELTS Listening IELTS Reading T2-Metaphor Layering-R IELTS Speaking T6-Heuristic-R	_	.31	55	16	.60	56	16	.65	111	32	.62
3	EGMC	4	T1-Phrasal Verbs-P OOPT Listening T1-Phrasal Verbs-R OOPT Use of English	_	.60	56	8	.31	56	11	.61	112	19	.46
4	EIMP	2	T6-Heuristic-P T7-Feelings-P	.36[b]	_	56	8	.54	56	8	.63	112	16	.58
5	ETVA	3	T4-Topic/Vehicle-R T5-Topic Transition-R T3-Vehicle Acceptability-R	_	.56	55	18	.69	55	18	.80	110	36	.73
6	EMLP	3	T8-Idiom Extension-P T9-Metaphor Continuation-P T8-Idiom Extension-R	_	.68	56	15	.85	56	14	.82	112	29	.83

Note. Code: (F1) EVS = English Vocabulary Size; (F2) EGC = English General Comprehension; (F3) EGMC = English Grammatical Metaphoric Competence; (F4) EIMP = English Illocutionary Metaphor Production; (F5) ETVA = English Topic/Vehicle Acceptability; (F6) EMLP = English Metaphor Language Play.
[a] Calculated as follows: *N* participants and *K* items = the sum of group 1 and group 2 values; alpha = the mean of group 1 and group 2 alpha values.
[b] Statistically significant, $p < .001$.

Definitions

α	Cronbach's alpha
BLC and HLC	Basic Language Cognition / Higher Language Cognition
BNC-BYU	British National Corpus-Brigham Young University
CFA	Confirmatory Factor Analysis
CI(s)	Confidence interval(s)
CMT	Conceptual metaphor theory
DMT	Deliberate metaphor theory
EFA	Exploratory Factor Analysis
EFL	English as a foreign language
ELF	English as a lingua franca
IQR	Interquartile range
K	The number of items
Kw	Weighted kappa
L1	First (native) language
L2	Second or foreign language
L3	A second or foreign language that is known less well than another one
LED	Longman English Dictionary
M	Mean
MC	Metaphoric competence
Mdn	Median
MED	Macmillan English Dictionary
MIP	Metaphor Identification Procedure
MIPVU	Metaphor Identification Procedure Vrije Universiteit (Amsterdam)
MRW(s)	Metaphor-related word(s)
NLS	Nativelike selection
NNS(s)	Non-native speaker(s)
NS(s)	Native speaker(s)
OED	Oxford English Dictionary
-P	-Productive (test)
PCA	Principal Components Analysis
Po	Percentage agreement
-R	-Receptive (test)
SD	Standard deviation
SEM	Structural Equation Modelling (family members include CFA, EFA and PCA)
SLA	Second language acquisition
VIP	Vehicle Identification Procedure
VU AMC	Vrije Universiteit Amsterdam Metaphor Corpus
WIDLII	When in doubt, leave it in (a metaphor code)

References

Aiken, L. R. (2003). *Psychological testing and assessment*. Boston: Allyn and Bacon.

Alderson, J. C. (1995). *Diagnosing foreign language proficiency*. London: Continuum.

Aleshtar, M. H., & Dowlatabadi, H. (2014). Metaphoric competence and language proficiency in the same boat. *Procedia - Social and Behavioral Sciences, 98*, 1895-1904.

Alexander, R. J. (1983). *Metaphors, connotations, allusions: thoughts on the languageculture connexion in learning English as a foreign language.* Paper presented at the L.A.U.T. papers, University of Trier.

Allami, H., & Aghajari, J. (2014). Pragmatic knowledge of assessment in listening sections of IELTS tests. *Theory and Practice in Language Studies, 4*(2), 332-340.

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

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Huston (Ed.), *Advances in reading/language research* (Vol. 2, pp. 231-256). Greenwich CT: JAI Press.

Andringa, S. (2014). The use of native speaker norms in critical period hypothesis research. *Studies in Second Language Acquisition, 36*(3), 565-596.

Anglim, J. (2016). Stack exchange. Retrieved from https://stats.stackexchange.com/questions/140460/how-can-i-calculate-the-95-confidence-interval-of-an-effect-size-if-i-have-the

Azuma, M. (2005). *Metaphorical competence in an EFL context*. Tokyo: Toshindo.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: University Press.

Bachman, L. F., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449-465.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford, New York: Oxford University Press.

Bayram, F., Rothman, J., Iversond, M., Kupisch, T., Miller, D., Puig-Mayenco, E., & Westergaard, M. (2017). Differences in use without deficiencies in competence: passives in the Turkish and German of Turkish heritage speakers in Germany. *International Journal of Bilingual Education and Bilingualism*, 1-21. doi:10.1080/13670050.2017.1324403

Beaty, R. E., & Silvia, P. J. (2013). Metaphorically speaking: cognitive abilities and the production of figurative language. *Memory & Cognition, 41*(2), 255-267.

Becker, L. (2000). Effect size calculators. Retrieved from http://www.uccs.edu/~lbecker/

Bell, N. (2005). Exploring L2 language play as an aid to SLL: A case study of humor in NS-NNS interaction. *Applied Linguistics, 26*(2), 192-218.

Bell, N. (2012). Formulaic language, creativity, and language play in a second language. *Annual Review of Applied Linguistics, 32*, 189-205.

Belz, J., & Reinhardt, J. (2004). Aspects of advanced foreign language proficiency: Internet-mediated German language play. *International Journal of Applied Linguistics, 14*(3), 324-362.

Blommaert, J. (2010). *The sociolinguistics of globalization.* Cambridge: Cambridge University Press.

Boerger, M. (2005). Variations in figurative language use as a function of mode of communication. *Journal of Psycholinguistic Research, 34*(1), 31-49.

Boers, F. (2000). Metaphor awareness and vocabulary retention. *Applied Linguistics, 21*(4), 553-571.

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

Boers, F., Demecheleer, M., & Eyckmans, J. (2004). Etymological elaboration as a strategy for learning figurative idioms. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 53-78). Amsterdam/Philadelphia: John Benjamins.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, H. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10*(3), 254-261.

Boers, F., & Lindstrimberg, F. (2012). Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics, 32*, 83-110.

Boers, F., & Littlemore, J. (2000). Cognitive style variables in participants' explanations of conceptual metaphors. *Metaphor and Symbol, 15*(3), 177-187.

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review, 112*(1), 193-216.

British Council: Common European Framework Equivalencies. (2017). Retrieved from http://takeielts.britishcouncil.org/find-out-about-results/understand-your-ielts-scores/common-european-framework-equivalencies

Brown, J. D. (2009). Statistics Corner. Questions and answers about language testing statistics: Choosing the right number of components or factors in PCA and EFA. *JALT Testing & Evaluation SIG Newsletter, 13*(2), 19-23.

Cameron, L. (2003). *Metaphor in educational discourse*. London: Continuum.

Cameron, L., & Deignan, A. (2006). The emergence of metaphor in discourse. *Applied Linguistics, 27*(4), 671-690.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Candlin, C. N. (1986). Explaining communicative competence limits of testability. In C. W. Stansfield (Ed.), *Toward communicative competence testing: Proceedings of the second TOEFL invitational conference* (pp. 38-57). Princeton, NJ: Educational Testing Service.

Carey, M., & Harrington, M. (2009). The on-line Yes/No test as a placement tool. *System, 37*(4), 614-627.

Carter, R., & McCarthy, M. (2004). Talking, creating: Interactional language, creativity and context. *Applied Linguistics, 25*(1), 62-88.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Chen, Y.-C. (2011). *The effects of teaching EFL learners metaphor and metonymy: With reference to emotion expressions* (Unpublished PhD dissertation). National Chengchi University, Taiwan.

Chen, Y.-C., & Lai, H. L. (2015). Developing EFL learners' metaphoric competence through cognitive-oriented methods. *Iral-International Review Of Applied Linguistics In Language Teaching, 53*(4), 415-438.

Chen, Y.-C., Lin, C.-Y., & Lin, S. (2014). EFL learners' cognitive styles as a factor in the development of metaphoric competence. *Journal of Language Teaching and Research, 5*(3), 698-707.

Chiappe, D. L., & Chiappe, P. (2007). The role of working memory in metaphor production and comprehension. *Journal of Memory and Language, 56*(2), 172-188.

Cogo, A. (2011). English as a Lingua Franca: concepts, use, and implications. *ELT Journal, 66*(1), 97-105.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). London: Routledge.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillside, NJ: Lawrence Erlbaum.

Cook, G. (1997). Language play, language learning. *ELT Journal, 51*(3), 224.

Cook, G. (2000). *Language play, language learning*. Oxford: Oxford University Press.

Corts, D. P., & Meyers, K. (2002). Conceptual clusters in figurative language production. *Journal of Psycholinguistic Research, 31*(4), 391-408.

Craik, F. I. M., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Crystal, D. (1998). *Language play*. London, UK: Penguin.

Dagut, M., & Laufer, B. (1985). Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition, 7*(1), 73-79.

Dai, Z., & Ding, Y. (2010). Effectiveness of text memorization in EFL learning of Chinese students. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 71-87). New York, NY: Continuum.

Danesi, M. (1992). Metaphorical competence in second language acquisition and second language teaching: The neglected dimension. In J. E. Alatis (Ed.), Georgetown University Round Table on Languages and Linguistics (pp. 489-500). Washington DC.

Danesi, M. (1995). Learning and teaching languages: The role of 'conceptual fluency'. *International Journal of Applied Linguistics, 5*(1), 3-20.

Davies, M. (2004-). British National Corpus (BYU-BNC). http://corpus.byu.edu/bnc/

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

Deignan, A., Gabrys, D., & Solska, A. (1997). Teaching English metaphors using cross-linguistic awareness-raising activities. *ELT Journal, 51*(4), 352-352.

Del Re, A. C. (2013). compute.es: Compute Effect Sizes. R package version 0.2-2.

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 14*(20), 1-11.

Doiz, A., & Elizari, C. (2013). Metaphoric competence and the acquisition of figurative vocabulary in foreign language learning. *Estudios de Lingüística Inglesa Aplicada, 13*, 47-82.

Drew, P., & Holt, E. (1998). Figures of speech: Figurative expressions and the management of topic transition in conversation. *Language in Society, 27*(4), 495-522.

Ellis, N. (2002). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition, 24*, 297-339.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Feng, G. C. (2014). Intercoder reliability indices: disuse, misuse, and abuse. *Quality and Quantity, 48*, 1803-1815.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London: Sage.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: Sage.

Fitzpatrick, T. (2007). Word association patterns: unpacking the assumptions. *International Journal of Applied Linguistics, 17*(3), 319-331.

Foster, P., Bolibaugh, C., & Kotula, A. (2014). Knowledge of nativelike selections in an L2: The Influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition, 36*, 101-132.

Foster, P., & Tavakoli, P. (2009). Lexical diversity and lexical selection: A comparison of native and non-native speaker performance. *Language Learning, 59*(4), 866-896.

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly, 41*(2), 339-359.

Gardner, H., Kirchner, M., Winner, E., & Perkins, D. (1975). Children's metaphoric productions and preferences. *Journal of Child Language, 2*, 125-141.

Gass, S., & Mackey, A. (2002). Frequency effects and second language acquisition: A complex picture? *Studies in Second Language Acquisition, 24*, 249-260.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition, 8*(2), 149-156.

Gibbs, R. W. (2011). Evaluating Conceptual Metaphor Theory. *Discourse Processes, 48*(8), 529-562. doi:10.1080/0163853X.2011.606103

Gibbs, R. W., & Chen, E. (2016). Taking metaphor studies back to the stone age: A reply to Xu, Zhang, and Wu (2016). *Intercultural Pragmatics, 14*(1), 117-124.

Giora, R. (1999). On the priority of salient meanings: Studies of literal and figurative language. *Journal of Pragmatics, 31*(7), 919-929.

Giora, R. (2003). *On our mind: salience, context, and figurative language* Oxford: Oxford University Press.

Glucksberg, S., & Haught, C. (2006). On the relation between metaphor and simile: When comparison fails. *Mind and Language, 21*(3), 360-378.

Goo, J., & Mackey, A. (2013). The case against the case against recasts. *Studies in Second Language Acquisition, 35*(1), 127-165.

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

GOV.UK: Home > Visas and immigration > Student visas > Tier 4 (General) student visa > 3. Knowledge of English. (2017). Retrieved from https://www.gov.uk/tier-4-general-visa/knowledge-of-english

Grady, J. E. (1997). Theories are buildings revisited. *Cognitive Linguistics, 8*(4), 267-290.

Grady, J. E. (1999). A typology of motivation for metaphor: Correlations vs. resemblances. In R. W. Gibbs & G. Steen (Eds.), *Metaphor in cognitive linguistics*. Amsterdam: John Benjamins.

Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W.-J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational And Psychological Measurement, 72*, 357-374.

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

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430-450.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*, 149-161.

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective – Challenges and potential solutions. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*: Eurosla Monographs Series, 2.

Hair, J., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, N. J.: Lawrence Erlbaum Associates.

Harley, B., Cummins, J., Swain, M., & Allen, P. (1990). The nature of language proficiency. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of second language proficiency* (pp. 7-38). Cambridge, UK: Cambridge University Press.

Harshman, R. A., & Reddon, J. R. (1983). *Determining the number of factors by comparing real with random data: A serious flaw and some possible corrections.* Paper presented at the Proceedings of the Classification Society of North America at Philadelphia, Philadelphia, USA.

Hashemian, M., & Nezhad, M. R. T. (2007). The development of conceptual fluency & metaphorical competence in L2 learners. *Linguistik Online, 30*(1), 41-56.

Haught, C. (2013). A tale of two tropes: How metaphor and simile differ. *Metaphor and Symbol, 24*(8), 254-274.

Henriksen, B. (2008). Declarative lexical knowledge. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Vocabulary and writing in a first and second language* (pp. 22-66). Basingstoke, UK: Palgrave Macmillan.

Hofmann, R. J. (1978). Complexity and simplicity as objective indices descriptive of factor solutions. *Multivariate Behavioral Research, 13*, 247-250.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hsu, J.-Y., & Chiu, C.-Y. (2008). Lexical collocations and their relation to speaking proficiency of college EFL learners in Taiwan. *Asian EFL Journal, 10*, 181-204.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Score on a Yes-No vocabulary test: Correction for guessing and response style. *Language Testing, 19*(3), 227-245.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers:Anagenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*, 229-249.

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition, 15*, 422-433.

Hulstijn, J. H., & Marchena, E. (1989). Avoidance: Grammatical or semantic causes? Studies in Second Language Acquisition. *Studies in Second Language Acquisition, 11*, 241-255.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: selected readings* (pp. 269-293). Harmondsworth: Penguin.

IELTS: Homepage. (2017). Retrieved from https://www.ielts.org/what-is-ielts/ielts-introduction

Ilić, M. (2014). Collective narrative: The narrative on Croatian language from academic to far-right discourses in Serbia. *Journal of Comparative Research in Anthropology and Sociology, 5*(1), 49-73.

Jenkins, J. (2006). Points of view and blind spots: ELF and SLA. *International Journal of Applied Linguistics, 16*(2), 137-162.

Johnson, J., & Pascual-Leone-J. (1989). Developmental levels of processing in metaphor interpretation. *Journal of Experimental Child Psychology, 48*(1), 1-41.

Johnson, J., & Rosano, T. (1993). Relation of cognitive style to metaphor interpretation. *Applied Psycholinguistics, 14*, 159-175.

Joliffe, I. T. (1972). Discarding variables in a principle component analysis. I: Artificial data. *Applied Statistics, 21*(2), 160-173.

Joliffe, I. T. (1986). *Principal Component Analysis*. New York: Springer.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational And Psychological Measurement, 20*, 141-151.

Kathpalia, S. S., & Carmel, H. L. H. (2011). Metaphorical competence in ESL student writing. *RELC Journal, 42*(3), 273-290.

Katz, A., Paivio, A., Marschark, M., & Clark, J. (1988). Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbolic Activity, 3*(4), 191-214.

Katz, I. (2013). Paxman interviews Boris Johnson (30 Sept. 2013). Retrieved from https://www.youtube.com/watch?v=Z_WFNB3tQz4

Kellerman, E. (1983). Now you see it, now you don't. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 112-134). Rowley, MA: Newbury House.

Kenny, D. A. (2015). Measuring model fit. Retrieved from http://davidakenny.net/cm/fit.htm

Keshavarz, M. H., & Salimi, H. (2007). Collocational competence and cloze test performance: A study of Iranian EFL learners. *International Journal of Applied Linguistics, 17*, 81-92.

Khamis, H., & Kepler, M. (2010). Sample size in multiple regression: 20 + 5k. *Journal of Applied Statistical Science, 17*, 505-517.

Kim, Y.-H., & Kellogg, D. (2007). Rules out of roles: Differences in play language and their developmental significance. *Applied Linguistics, 28*, 25-45.

Klein-Braley, C. (1985). A cloze-up on the C-test: A study in the construct validation of authentic tests. *Language Testing, 2*, 75-104.

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

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). London: Routledge.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: SAGE Publications.

Kövecses, Z. (2003). Language, figurative thought, and cross-cultural comparison. *Metaphor and Symbol, 18*(4), 311-320.

Kövecses, Z. (2010). *Metaphor: A practical introduction* (2nd ed.). Oxford: Oxford University Press.

Kövecses, Z., & Szabó, P. (1996). Idioms: A view from cognitive semantics. *Applied Linguistics, 17*(3), 326-355.

Krashen, S. (1977). Some issues relating to the monitor model. In H. Brown, C. Yorio, & R. Crymes (Eds.), *Teaching and learning English as a second language: Some trends in research and practice* (pp. 144-148). Washington, DC: TESOL.

Kupisch, T., & Rothman, J. (2016). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *International Journal of Bilingualism*, 1-19. doi:10.1177/1367006916654355

Kurtyka, A. (2001). Teaching English phrasal verbs: A cognitive approach. In M. Pütz & S. Niemeier (Eds.), *Applied cognitive linguistics II: Language pedagogy* (pp. 29-54). Berlin, New York: Mouton De Gruyter.

Lado, R. (1961). *Language testing*. New York, NY: McGraw-Hill.

LaFlair, G. T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t-tests, and anovas. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46-77). New York: Routledge.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lakoff, G., & Turner, M. (1989). *More than cool reason*. Chicago, Ilinois: University of Chicago Press.

Laufer, B. (1997). What's in a word that makes it hard or easy? Intralexical factors affecting vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 140-155). Cambridge: Cambridge University Press.

Laufer, B. (2005). Lexical frequency profiles: From Monte Carlo to the real world: A response to Meara (2005). *Applied Linguistics, 26*(4), 582-588.

Laufer, B., & Eliasson, S. (1993). What causes avoidance in L2 learning: L1-L2 difference, L1-L2 similarity, or L2 complexity? *Studies in Second Language Acquisition, 15*, 35-48.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.

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

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33-51.

Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing, 30*(1), 99-123.

Lefford, A. (1946). The influence of emotional subject matter on logical reasoning. *Journal of General Psychology, 34*, 127-151.

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Leow, R. P. (2000). A study of the role of awareness in foreign language behaviour. *Studies in Second Language Acquisition, 22*, 557-584.

Levorato, M. C., & Cacciari, C. (2002). The creation of new figurative expressions: psycholinguistic evidence in Italian children, adolescents and adults. *Journal of Child Language, 29*(1), 127-150.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning, 54*(2), 193-226.

Littlemore, J. (2001). Metaphoric competence: A language learning strength of students with a holistic cognitive style? *TESOL Quarterly, 35*(3), 459-491.

Littlemore, J. (2004). Item-based and cognitive-style-based variation in students' abilities to use metaphoric extension strategies. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos/Journal of the European Association of Languages for Specific Purposes (AELFE), 7*, 5-31.

Littlemore, J. (2005). Figurative thought and the teaching of languages for specific purposes. In E. Hernandez & L. Sierra (Eds.), *Lenguas para fines especificos (VIII) investigacion y ensenanza* (pp. 25-35). Madrid: University de Alcala.

Littlemore, J., & Low, G. D. (2006a). *Figurative thinking and foreign language learning*. Basingstoke: Palgrave Macmillan.

Littlemore, J., & Low, G. D. (2006b). Metaphoric competence, second language learning, and communicative language ability. *Applied Linguistics, 27*(2), 268-294.

Liu, X., & Xiao, G. (2013). A comparative study of emotion metaphors between English and Chinese. *Theory and Practice in Language Studies, 3*(1), 155-162.

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principle components analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 182-212). New York: Routledge.

Loewen, S., Lavolette, E., Spino, L., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly, 48*(2), 360-388.

Loewen, S., Li, S., Fei, F., Thompson, A., Nakatsukasa, K., Ahn, S., & Chen, X. (2009). Second language learners' beliefs about grammar instruction and error correction. *The Modern Language Journal, 93*(1), 91-104.

Low, G. D. (1988). On teaching metaphor. *Applied Linguistics, 9*(2), 125-147.

Lowry, R. (2017). VasserStats: Website for statistical computation, significance of the difference between two correlation coefficients. Retrieved from http://vassarstats.net/rdiff.html

Luk, Z. P., & Shirai, Y. (2009). Is the acquisition order of grammatcial morphemes imperveous to L1 knowledge? Evidence from the acquisition of plural s-, articles and possessive 's. *Language Learning, 59*(4), 721-754.

Lv, Z., & Zhang, Y. (2012). Universality and variation of conceptual metaphor of love in Chinese and English. *Theory and Practice in Language Studies, 2*(2), 355-359.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching, 46*(1), 1-40.

MacArthur, F. (2010). Metaphorical competence in EFL: Where are we and where should we be going? A view from the language classroom In J. Littlemore & C. Juchem-Grundmann (Eds.), *Applied cognitive linguistics in second language learning and teaching. AILA review* (Vol. 23, pp. 155-173).

Maccallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99.

Macmillan English Dictionary. (2017). Macmillan Publishers Limited.

Mashal, N., & Kasirer, A. (2011). Thinking maps enhance metaphoric competence in children with autism and learning disabilities. *Research in Developmental Disabilities, 32*(6), 2045-2054.

Mashal, N., & Kasirer, A. (2012). Principal component analysis study of visual and verbal metaphoric comprehension in children with autism and learning disabilities. *Research in Developmental Disabilities, 33*, 274-282.

McElree, B., & Nordlie, J. (1999). Literal and figurative interpretations are computed in equal time. *Psychonomic Bulletin & Review, 6*(3), 486-494.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt & M. M. (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 109-121). Cambridge: Cambridge University Press.

Meara, P. (2004). Modelling vocabulary loss. *Applied Linguistics, 25*(2), 137-155.

Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics, 26*(1), 32-47.

Meara, P. (2006). Emergent properties of multilingual lexicons. *Applied Linguistics, 27*(4), 620-644.

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

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

Meara, P., & Miralpeix, I. (2015). V_YesNo v1.0. Retrieved from www.lognostics.co.uk/

Meara, P., & Wolter, B. (2004). V_Links: Beyond vocabulary depth. In D. Albrechtsen & B. H. K. Haastrup (Eds.), *Angles on the English-speaking world 4* (pp. 85-96). Copenhagen: Museum Tusculanum Press.

Melka, F. (1997). Receptive versus productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 84-102). New York: Cambridge University Press.

Mitchell, R., Myles, R., & Marsden, E. (2013). *Second language learning theories*. Abingdon, UK: Routledge.

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Clarendon Press.

Müller, C. (2008). *Metaphors dead and alive, sleeping and waking: A dynamic view*. Chicago, Ilinois: University of Chicago Press.

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes. *Studies in Second Language Acquisition, 38*, 365-401.

Murphy, G. (1996). On metaphoric representations. *Cognition, 60*, 173-204.

Myles, F., & Cordier, C. (2017). Formulaic sequence (FS) cannot be an umbrella term in SLA. *Studies in Second Language Acquisition, 39*, 3-28.

Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning, 48*, 323-363.

Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2: A basis for creative construction? *Studies in Second Language Acquisition, 21*, 49-80.

Nacey, S. (2013). *Metaphors in learner English*. Amsterdam: John Benjamins.

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

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59-82.

Nation, I. S. P., & Beglar, D. (2007). vocabulary size test. *The Language Teacher, 31*, 9-17.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Neagu, M. (2007). English verb particles and their acquisition: A cognitive approach. *RESLA, 20*(121-138).

Noro, T. (2002). The roles of depth and breadth of vocabulary knowledge in reading comprehension in EFL. *ARELE, 13*, 71-80.

Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research, 34*(2), 257-271.

Norris, J. M., & Ortega, L. (2006). *Synthezing research on language learning and teaching*. Amsterdam: John Benjamins.

NourMohamadi, E. (2010). *Conceptual metaphor and the acquisition of English metaphorical competence by Persian English majors: A cognitive linguistic approach* (Unpublished PhD dissertation). Allame University, Iran.

Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes, 18*, 161-175.

Oxford Reference Dictionary. (2017). Oxford University Press.

Pallant, J. (2013). *SPSS survival manual: A step-by-step guide to data analysis using SPSS* (4th ed.). Maidenhead: McGraw-Hill /Open University Press.

Palmberg, R. (1987). Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition, 9*, 202-221.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130-149.

Patil, V. H., Singh, S. N., Mishra, S., & Donavan, T., D. (2008). Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one' criterion. *Journal of Business Research, 61*, 162-170.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing, 29*(4), 489-509.

Penfield, W., & Roberts , L. (1959). *Speech and brain mechanisms*. Princeton, NJ: Princeton University Press.

Pettersson, E., & Turkheimer, E. (2010). Item selection, evaluation, and simple structure in personality data. *Journal of research in personality, 44*(4), 407-420.

Philip, G. (2010). "Drugs, traffic, and many other dirty interests": Metaphor and the language learner. In G. Low, Z. Todd, A. Deignan, & L. Cameron (Eds.), *Researching and Applying Metaphor in the Real World* (pp. 63-80). Amsterdam; Philadelphia: John Benjamins.

Pickens, J., & Pollio, H. R. (1979). Patterns of figurative language competence in adult speakers. *Psychological Research, 40*(3), 299-313.

Pilcher, N., & Richards, K. (2016). The paradigmatic hearts of subjects which their 'English' flows through. *Higher Education Research & Development, 35*(5), 997-1010.

Pilcher, N., & Richards, K. (2017). Challenging the power invested in the International English Language Testing System (IELTS): Why determining 'English' preparedness needs to be undertaken within the subject context. *Power and Education, 9*(1), 3-17.

Pitzl, M.-L. (2009). 'We should not wake up any dogs': Idiom and metaphor in ELF. In A. Mauranen & E. Ranta (Eds.), *English as a lingua franca: Studies and findings* (pp. 299-322). Newcastle upon Tyne: Cambridge Scholars.

Pitzl, M.-L. (2016). World Englishes and creative idioms in English as a lingua franca. *World Englishes, 35*(2), 293-309.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition, 35*(4), 655-687.

Plonsky, L., & Derrick, D. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538-553.

Plonsky, L., Egbert, J., & LaFlair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics, 36*(5), 591-610.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning, 65*(Supplement 1), 9-36.

Plonsky, L., & Oswald, F. L. (2014). How big Is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 879-912.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition, 39*(3), 579-592.

Pollio, H. R. (1977). *Personal cognition and metaphoric style*. Paper presented at the American Psychological Association, San Francisco.

Pollio, H. R., Barlow, J. M., Fine, H. J., & Pollio, M. R. (1977). *Psychology and the poetics of growth: Figurative language in psychology, psychotherapy, and education*. Hillside, NJ: Lawrence Erlbaum Associates.

Pollio, H. R., & Burns, B. (1977). The anomaly of anomaly. *Journal of Psycholinguistic Research, 6*(3), 247-260.

Pollio, H. R., & Smith, M. (1979). Sense and nonsense in thinking about anomaly and metaphor research. *Bulletin of the psychonomic society, 13*(5), 323-326.

Pollio, H. R., & Smith, M. (1980). Metaphoric competence and complex human problem solving. In R. P. Honeck & R. R. Hoffman (Eds.), *Cognition and figurative language* (pp. 365-392). Hillside, NJ: Lawrence Erlbaum.

Pollio, M. R., & Pollio, H. R. (1979a). The comprehension of figurative language in children. *Journal of Child Language, 6*, 111-120.

Pollio, M. R., & Pollio, H. R. (1979b). A test of metaphoric comprehension and some preliminary developmental data. *Journal of Child Language, 6*(1), 111-120.

Pollitt, A. (2016). The Oxford Online Placement Test: The meaning of OOPT scores. Retrieved from https://www.oxfordenglishtesting.com/defaultmr.aspx?id=3048

Pragglejaz. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol, 22*(1), 1-39.

Prodromou, L. (2003). The idiomatic paradox and English as a lingua franca. *Modern English Teacher, 12*(1), 22-29.

Purpura, J. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Purpura, J. (2009). The Oxford Online Placement Test: What does it measure and how? Retrieved from https://www.oxfordenglishtesting.com/defaultmr.aspx?id=3048

Qi, G.-Y. (2016). The importance of English in primary school education in China: perceptions of students. *Multilingual Education, 6*(1), 1-18.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*, 282–307.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355-371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41-60). Mahwah, NJ: Lawrence Erlbaum.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Reijnierse, W. G., Burgers, C., Krennmayr, T., & Steen, G. (under review). DMIP: A method for identifying potentially deliberate metaphor in language use.

Revelle, W. (2017). Procedures for psychological, psychometric, and personality research (Version 1.6.12). Retrieved from http://personality-project.org/r/psych-manual.pdf

Richard, J.-P. J. (2011). Does size matter? The relationship between vocabulary breadth and depth. *Sophia International Review, 33*, 107-120.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*, 77-89.

Richards, K., & Pilcher, N. (2016). An individual subjectivist critique of the use of corpus linguistics to inform pedagogical materials. *Dialogic Pedagogy Journal, 4*.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*, 135-147.

Ringbom, H. (1987). *The role of the first language in foreign language learning*. Clevedon, England: Multilingual Matters.

Roberts, L., & Felser, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics, 32*(2), 299-331.

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.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24*, 282-292.

Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language Learning, 43*, 313-344.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review, 11*, 11-26.

Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing, 16*, 189-216.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913-951.

Schmitt, N., & 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., 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.

Searle, J. (1993). Metaphor. In A. Ortony (Ed.), *Metaphor and thought* (2nd ed., pp. 92-123). Cambridge: Cambridge University Press.

Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics, 11*(2), 153-158.

Seidlhofer, B. (2005). English as a lingua franca. In A. S. Hornby (Ed.), *Oxford advanced learner's dictionary of current English*. Oxford: Oxford University Press.

Seidlhofer, B., Breiteneder, A., Klimpfinger, T., Majewski, S., Osimk-Teasdale, R., Pitzl, M.-L., & Radeka, M. (2013). *VOICE: The Vienna-Oxford International Corpus of English*.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics, 10*, 209-231.

Semino, E., & Masci, M. (1996). Politics is football: Metaphor in the discourse of Silvio Berlusconi in Italy. *Orbis, 7*(2), 243-270.

Sewell, A. (2013). English as a lingua franca: Ontology and ideology. *ELT Journal, 67*(1), 3-10.

Silvia, P. J., & Beaty, R. E. (2012). Making creative metaphors: The importance of fluid intelligence for creative thought. *Intelligence, 40*(4), 343-351.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532.

Steen, G. (2008). The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor and Symbol, 23*(4), 213-241.

Steen, G. (2011a). From three dimensions to five steps: The value of deliberate metaphor. *Metaphorik.de, 21*, 83-110.

Steen, G. (2011b). What does 'really deliberate' really mean? More thoughts on metaphor and consciousness. *Metaphor & the Social World, 1*, 53-56.

Steen, G. (2013). Deliberate metaphor affords conscious metaphorical cognition. *Journal of Cognitive Semiotics, 5*, 179-197.

Steen, G. (2015). Developing, testing and interpreting Deliberate Metaphor Theory. *Journal of Pragmatics, 90*, 67-72.

Steen, G. (2016). Mixed metaphor is a question of deliberateness. In R. W. Gibbs (Ed.), *Mixing metaphor* (pp. 113–132). Amsterdam: John Benjamins.

Steen, G. (2017). Deliberate Metaphor Theory: Basic assumptions, main tenets, urgent issues. *Intercultural Pragmatics, 14*(1), 1-24.

Steen, G., Dorst, A. G., Berinke Herrmann, J., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification : from MIP to MIPVU*. Amsterdam: John Benjamins.

Steen, G., Reijnierse, W. G., & Burgers, C. (2014). When do natural language metaphors influence reasoning? A follow-up study to Thibodeau and Boroditsky (2013). *PLoS ONE, 9*(12), e113536. doi:10.1371/journal.pone.0113536

Steinberg, D. (1970). Analyticity, amphigory, and the semantic interpretation of sentences. *Journal of Verbal Learning and Verbal Behavior, 9*, 37-51.

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: does the type of target language influence the association? *IRAL - International Review of Applied Linguistics in Language Teaching, 49*(4), 321-343.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillside, NJ: Erlbaum.

Stumberg, D. (1928). A study of poetic talent. *Journal of Experimental Psychology, 20*, 214-230.

Su, L. I.-W. (2002). What can metaphors tell us about culture? *Language and Linguistics, 3*(3), 589-613.

Sullivan, P. (2000). Playfulness as mediation in communicative language teaching in a Vietnamese classroom. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 115-131). Oxford, UK: Oxford University Press.

Sung, C. C. M. (2012). *English pronunciation, identity and pedagogy: Exploring the perceptions of L2 speakers of English as a Lingua Franca in Hong Kong* (Unpublished PhD dissertation). Lancaster University, UK.

Sung, C. C. M. (2013). English as a Lingua Franca and English language teaching: A way forward. *ELT Journal, 67*(3), 350-353.

Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

Tabachnick, B. G., & Fidell, L. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson Education.

Taki, S., & Soghady, M. R. N. (2013). The role of L1 in L2 idiom comprehension. *Journal of Language Teaching and Research, 4*(4), 824.

Thompson, B., & Levitov, J. E. (1985). Using microcomputers to score and evaluate test items. *Collegiate Microcomputer, 3*, 163-168.

Thurstone, L. L. (1935). *The vectors of mind*: University of Chicago Press.

Tin, T. B. (2011). Language creativity and co-emergence of form and meaning in creative writing tasks. *Applied Linguistics, 32*, 215-235.

Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology, 34*(4), 414-424.

Torrance, E. P. (1974). *Torrance tests of creative thinking*. Lexington, Mass: Ginn & Co.

Tourangeau, R., & Sternberg, R. (1981). Aptness in metaphor. *Cognitive Psychology, 13*, 27-25.

Trahey, M., & White, L. (1993). Positive evidence and premeption in the second language classroom. *Studies in Second Language Aquisition, 15*, 181-204.

Tseng, W.-T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning, 58*(2), 357-400.

ttnphns. (2012). Interpreting discrepancies between R and SPSS with exploratory factor analysis. Retrieved from https://stats.stackexchange.com/questions/24781/interpreting-discrepancies-between-r-and-spss-with-exploratory-factor-analysis

Turner, N. E. (1998). The effect of common variance and structure pattern on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational And Psychological Measurement, 58*, 541-568.

UK Council for International Student Affairs: International (non-UK) students in UK HE in 2015-16. (2017). Retrieved from https://institutions.ukcisa.org.uk/Info-for-universities-colleges--schools/Policy-research--statistics/Research--statistics/International-students-in-UK-HE/

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217-234.

Vervaeke, J., & Kennedy, J. (1996). Metaphors in language and thought: Disproof and multiple meanings. *Metaphor and Symbolic Activity, 11*, 273-284.

Wang, L.-C., & Hyun, E. (2009). A study of sociolinguistic characteristics of Taiwan children's peer talk in a Mandarin English-speaking preschool. *Journal of Early Childhood Research, 7*, 3-26.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*, 33-52.

Webb, S., & Rodgers, M. P. H. (2009a). The lexical coverage of movies. *Applied Linguistics, 30*(3), 407-427.

Webb, S., & Rodgers, M. P. H. (2009b). The vocabulary demands of television programs. *Language Learning, 59*(2), 335-366.

Wesche, M. B., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review, 53*(1), 13-40.

Williams, J. (2001). The effectiveness of spontaneous attention to form. *System, 25*, 129-140.

Winner, E., & Gardner, R. (1977). Sensitivity to metaphor in organic patients. *Brain, 100*, 719-727.

Winner, E., Rosentiel, A. K., & Gardner, H. (1976). The development of metaphoric understanding. *Developmental psychology, 12*(4), 289.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford, UK: Oxford University Press.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics, 32*, 231-254.

Yasuda, S. (2012). Learning phrasal verbs through conceptual metaphors: A case of Japanese EFL learners. *TESOL Quarterly, 44*(2), 250-273.

Yu, N. (1998). *The contemporary theory of metaphor: A perspective from Chinese*. Amsterdam: John Benjamins.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal, 96*(4), 558-575.

Zhang, W. (2013). An exploration of "anger" metaphor translations based on cognitive equivalence hypothesis (CEH). *Theory and Practice in Language Studies, 3*(5), 790-796.

Zhao, Q., Yu, L., & Yang, Y. (2014). Correlation between receptive metaphoric competence and reading proficiency. *English Language Teaching, 7*(11), 168-181.

Žiberna, A. (2015). Interpreting discrepancies between R and SPSS with exploratory factor analysis. Retrieved from https://stats.stackexchange.com/questions/24781/interpreting-discrepancies-between-r-and-spss-with-exploratory-factor-analysis

Zientel, L. R., & Thompson, B. (2007). Applying the bootstrap to the multivariate case: Bootstrap component/factor analysis. *Behavior Research Methods, 39*(2), 318-325.

Zopluoglu, C. (2017a). An R routine to construct confidence intervals for sample eigenvalues. Retrieved from https://sites.education.miami.edu/zopluoglu/software-programs/

Zopluoglu, C. (2017b). An R routine to estimate standard errors for factor loadings in EFA. Retrieved from https://sites.education.miami.edu/zopluoglu/software-programs/