Accepted 'preprint' version.

Article accepted on 26 April 2021

Citation:

O'Reilly, D. & Marsden, E. (in press). Elicited metaphoric competence in a second language: A construct associated with vocabulary knowledge and general proficiency? *International Review of Applied Linguistics.*

Elicited metaphoric competence in a second language: A construct associated with vocabulary knowledge and general proficiency?

Abstract

The extent to which the ability to use metaphor in a second language ¬- metaphoric competence (MC) - relates to well-attested language proficiency components has implications both for understanding second language (L2) competence and for pedagogy. Building on previous enquiries (Azuma 2005) and extending a vocabulary size and depth research agenda (Qian 2002; Schmitt 2014) to the realm of MC, the present study sought to disentangle the relationships between six elicited MC constructs, reliably established by O'Reilly and Marsden (2021), two standardised L2 proficiency measures, and established vocabulary size and depth measures. With 108 Mandarin learners of L2 English, partial correlation analyses showed unique relationships between specific MC and proficiency measures, evidence of what these learners could do with metaphor at various proficiency levels, and how sparse references to metaphor in proficiency descriptors (e.g., CEFR) might be more precisely interpreted. Multiple regression analyses showed that Read's (1993, 1998) Word Associates Test, a vocabulary depth measure, was closely linked to all types of MC, particularly productive control (Henriksen 1999) and metaphor language play (O'Reilly and Marsden 2021). The findings point to the centrality of different (but related) types of associative thinking ability in metaphor use and language learning more generally (Carroll 1993; Littlemore 2001; 2002; 2008; Littlemore and Low 2006a). Future research implications and pedagogical reflections are provided.

1. Introduction

Metaphor, broadly "...a reclassification which involves: Treating X as if it were, in some ways, Y" (Low 1988: 126), is used by all first (L1) and second language (L2) speakers. For present purposes, we use the term *metaphor* to denote all types of figurative language, including metaphor, metonymy, simile, idiom, and other tropes that learners are likely to encounter. Metaphor is now known to form a large amount of day-to-day language, in English, some 17.5% of academic discourse, 15.3% of news texts, 10.8% of fiction, and 6.8% of spoken conversation (see Steen et al. 2010: 194-208). People use metaphor, both consciously and unconsciously, for talking about abstract entities, persuading others to a certain way of thinking, shifting blame, negotiating, explaining medical diagnoses, making jokes, maintaining relationships, and for a variety of other purposes. L2 learners, therefore, have much to gain by becoming proficient users of metaphor.

A distinction is often made between *linguistic* metaphors (metaphor in language) and underlying *conceptual* metaphors (metaphor in thought) that, to some extent, motivate and shape them (e.g., the conceptual metaphor HAPPINESS IS UP underlies *I feel high*, *raise our spirits*, and *cheer up*). At the linguistic level, the Topic (feeling happy) is conveyed by the Vehicle term (the actual words used, *I feel high*), indicating a conceptual mapping between source domain (UP) and target domain (HAPPINESS) (Lakoff and Johnson 1980).

Metaphoric competence (MC) then, can be regarded as the ability to use metaphorical language, ideas, and systems conventionally, creatively, strategically, and skilfully, and both its linguistic and conceptual facets are therefore relevant to the present study. To help understand the nature of L2 MC and its current and potential role in pedagogy, Low's (1988) and Littlemore and Low's (2006a; 2006b) longstanding and well-cited accounts offer detailed descriptions of metaphor-related skills and (sub)competences, and suggestions for developing these in the L2 classroom. Low (1988: 131), described (for example) the need for learners to develop knowledge of acceptable word and idea manipulation (e.g., "[that] one ordinarily says 'The river snaked (its way) through the jungle', but not 'The river was (like)/resembled a snake'"), and the ability to successfully continue a metaphor throughout a conversation, as L1 speakers are often expected to do. Extending this work, Littlemore and Low (2006a; 2006b) applied metaphor and figurative thinking to Bachman's (1990) model of *communicative competence*, describing (for example) the illocutionary/imaginative skill of playfully re-literalising idioms in acceptable ways (e.g., "I've been sitting on the fence so long my bottom is beginning to hurt", 2006a: 130), and the heuristic skill of using metaphor to provide ad-hoc explanations of things (e.g., *the brain is a computer*).

To operationalise and elicit Low/Littlemore's constructs, a large battery of MC tests has been developed and its reliability ascertained. Both short and long versions are now available at www.iris-database.org (Marsden, Mackey, and Plonsky 2016). In O'Reilly and Marsden (2021), the MC Test

Battery was administered to 112 L1 Mandarin L2 English learners and, to establish several scoring parameters, 31 L1 English speakers. Various statistical techniques were used to detect and remove rogue items and extreme cases, thus optimising item-within-test consistency (reliability) and achieving a closer measure of the intended MC constructs (validity). The test battery overcame methodological limitations of previous instrumentation, including the range and number of MC sub-components measured, the number of test items per construct, and the use of ordinal omega, a powerful alternative to Cronbach's alpha, to estimate test reliability. Scores for the 15 MC tests showed areas of relative ease and difficulty, and were used to computer overall/composite receptive and productive MC. Exploratory factor analysis (a hitherto unused technique in L2 MC research) revealed four latent/underlying variables in 15 MC tests, interpreted/labelled as *Productive Illocutionary MC*, *Metaphor Language Play*, *Topic/Vehicle Acceptability*, and *Grammatical MC*. However, the relationships between MC (as measured) and other aspects of language competence were not explored.

In the L2 MC literature, a few studies using elicitation methods provide tentative evidence that receptive and productive MC correlates positively with both vocabulary *size* and *depth* (Azuma 2005) and general proficiency (Aleshtar and Dowlatabadi 2014; Zhao et al. 2014), and that MC-skill aspects link to a holistic (rather than analytic) cognitive style (Littlemore 2001). However, there is a need to move from simple bivariate correlation analysis to techniques that offer more nuanced, controlled ways to explore interrelatedness such as partial correlation and multiple regression (Linck and Cunnings 2015; Plonsky and Oswald 2017). While regression is common in L2 vocabulary research (Schmitt 2010) and has enabled comparisons of cognitive-oriented and metaphor-mapping methods for fostering metaphor/metonymy recognition (Chen and Lai 2015), its usefulness for identifying how robust MC construct measures (e.g., the MC Test Battery in O'Reilly and Marsden 2021) relate to vocabulary knowledge remains unexplored.

These issues have important implications for language teachers and testers seeking research-based guidance; what can teachers expect their learners to be able to do with metaphor at various proficiency levels? Does MC link differently to different high-stakes proficiency measures? What kinds of MC might develop alongside a large and broad vocabulary and what does this reveal about the MC construct? The present study begins to address these gaps.

2. Literature review

2.1. L2 MC, a matter of skill or knowledge?

At the core of MC are skill-based abilities to see potential metaphorical and metonymic relationships between language (e.g., polysemous words) and ideas.[1] These include cognitive abilities to reason,

[1] We thank one anonymous reviewer in particular for helpful comments on this aspect of the paper.

analyse, and solve problems (fluid intelligence); make a wide range of connections for a given stimulus (associative thinking/fluency); spot and exploit partial similarities between concepts (analogical reasoning); and form mental images (e.g., Littlemore and Low 2006a; Littlemore 2001; 2002; 2008). Associative thinking, in particular, is considered central to appropriate metaphor interpretation and production (Carroll 1993). For example, an L2 learner who knows that 'hot' literally means *very high in temperature* may be able to retrieve information broadly associated with this source domain (*burning/flames*, *danger*, *energy* etc.) to work out the target metaphorical meaning in 'hot temper' (*anger*/*strong emotion*) and 'hot topic' (*exciting*/*interesting*/*live*), and experiment with producing their own metaphors/metaphor variations. Some associations will be semantic (e.g., 'hot'-*burn-pain*), others syntactic (e.g., 'hot'-*water-cold-air*), and others connected to the stimulus in different ways (e.g., phonologically, typological similarity). Research has shown that divergent thinkers, who use vaguer, associational search criteria are more successful with unknown L2 metaphor than convergent thinkers, who favour single solutions to problems via rigid, logical search criteria (Littlemore 2001; Littlemore and Low 2006a).

MC skill-based components are likely to help learners acquire its more knowledge-based components, which inevitably relate to vocabulary knowledge (Section 2.2) and general proficiency (Section 2.3). MC 'knowledge' denotes stored form-meaning link information (e.g., that 'hot' can mean *high temperature*, *strong emotion* etc.), reliably established in and retrievable from the learner's memory, developing through increased experience and education (crystallised intelligence). To know that 'hot temper' means *gets angry easily*, is to have previously encountered this form-meaning link, be aware of it and able to efficiently and accurately deploy this knowledge in real-time comprehension or production with increasing mastery.

While learners are often unaware that they misunderstand metaphor (Littlemore et al. 2011) they can be trained (with some success) to use associative thinking, analogical reasoning, and mental imagery strategies to decipher unknown metaphors (Littlemore 2002; 2004c). Contextual clues may impact on what sort of cognitive processing learners engage in (Littlemore 2004a); a more holistic cognition seems conducive to efficient metaphor meaning comprehension, since both processes involve loose analogical reasoning (Littlemore 2001). Cognitive style also links to the ability to use (but not preference for using) a particular strategy for interpreting unknown metaphor (Littlemore 2004b), and L2ers with a stronger image forming capacity appear better at processing novel metaphors those with a verbalising style, who favour use of surrounding discourse context (Littlemore 2008). Interestingly, with reasonable L2 proficiency, L1-L2 differences become negligible for the execution of various MC skills (e.g., novel metaphor production); observed with both between-groups (Johnson and Rosano 1993) and within-groups designs (Littlemore 2010). Knowledge components of MC, on the other hand, would intuitively have a stronger relationship with general proficiency (Section 2.3).

In skill acquisition theory (DeKeyser 2017a; 2017b; 2018), MC *knowledge* is loosely akin to declarative knowledge (i.e., form-meaning mapping knowledge, probably with explicit awareness) and MC *skills* to proceduralised knowledge (i.e., more rapid and accurate communicative knowledge execution). Skill acquisition theory posits that declarative knowledge allows learners to engage in target behaviour with rules, structures etc. which, with enough meaningful repetition, enables proceduralised knowledge to develop, and for certain L2 learners and language structures, automatised knowledge (McManus and Marsden 2018). Proceduralised and automatised knowledge tend to be skill-specific (production practice fosters production, less so comprehension; listening practice fosters listening etc.). However, despite a rich research literature, the type of knowledge learners are actually using (declarative, proceduralised, automotised) is routinely difficult to decipher (DeKeyser 2017a).

The above points suggest a complex relationship between fluid, MC skills and crystallised, MC knowledge. Carroll's (1993) three-stratum cognitive ability theory suggests they sit on the same second level stratum under a higher, general intelligence construct. However, recent work has criticised the stratum II composition and interpretation (Benson et al. 2018). While engaging in MC skills (e.g., associative thinking) can help learners acquire language knowledge, the *type* of knowledge and system learning that skill development might promote is uncertain. Has the learner who uses 'hot temper' now understood that 'hot' can be used in different ways? that temperature adjectives (or adjectives more generally?) can be used metaphorically? Or something else? Similarly, one does not either *know* or *not know* a metaphor; knowledge is multifaceted, gradual, and dynamic (Section 2.2). Furthermore, since skill execution necessarily draws on known information permanently stored in memory, skills *are*, in a sense, a form of knowledge (DeKeyser 2017a).

In the current study we are primarily interested in *what* MC test-takers were able to demonstrate (rather than *how* they did this) and how this relates to more established language constructs (but see O'Reilly 2017 for test-taker introspections during piloting). Participants had freedom to gain marks for deploying any skills or knowledge when completing the untimed MC tests (Section 3.2). While cognitive skills probably lie at the core of the MC construct, we use the term 'MC' where either knowledge and skill components *might* be at play, and the terms [MC] 'skill' or 'knowledge' where the distinction is more clear-cut. We return to these points in the discussion.

2.2. L2 MC and vocabulary

In her influential think-piece, Henriksen (1999) proposed three core dimensions of L2 lexical competence: (1) partial-to-precise levels of knowledge (e.g., L2-L1 translation, multiple-choice definition recognition); (2) depth/quality of knowledge, particularly relationships between words in the lexicon; and (3) receptive-to-productive use, comprising control and accessibility. The author highlighted the interrelationships between the three dimensions, arguing from the available literature

that more precise understanding accompanies a greater depth of knowledge, necessary for better control in vocabulary production.

Much empirical work on L1/L2 vocabulary knowledge focuses on one or two of Henriksen's dimensions. One key issue is the extent of conceptual difference between vocabulary size (the number of forms a learner knows at least something about) and depth (the quality of knowledge of words known). Most researchers maintain a meaningful distinction between these two constructs (e.g., Gyllstad 2013; Meara and Wolter 2004; Schmitt 2014) but some (e.g., Vermeer 2001) emphasise their intuitive and statistical relatedness, and data showing they are equally good indicators of learners' vocabulary knowledge, as evidence for conceptual unity. Synthesising these perspectives, Schmitt (2014) suggests the size-depth distinction partly depends on how constructs are operationalised and advises the use of precise terminology when theorising about vocabulary knowledge. Vocabulary depth, for example, may be conceptualised in terms of mastery of a single aspect such as polysemy, or multiple aspects such as synonyms, collocations, derivations etc. (see Section 3.2 for our vocabulary size and depth conceptualisations).

Precise terminology is also important when determining what a particular vocabulary test actually measures. To date, vocabulary size test formats have included word-definition matching (e.g., Vocabulary Levels Test, Nation 1983); yes/no word recognition/checklist tests (e.g., V_YesNo, Meara and Miralpeix 2015), and mutilated-word gap-fill to measure controlled production (Laufer and Nation 1999). Vocabulary depth formats have included self-rating and/or provide-a-synonym tasks (e.g., Vocabulary Knowledge Scale, Wesche and Paribakht 1996), paradigmatic/analytic/syntagmatic associate matching (Word Associates Test, Read 1993; 1998; hereafter WAT), and tasks to decipher and supply target words with appropriate affixation (e.g., *arm, disarm, armed*) from sentences containing gaps (e.g., 1K-Vocabulary Depth Test, Richard 2011).

Both vocabulary size and depth seem to contribute to overall L2 proficiency. While some studies show vocabulary size to be better at predicting L2 proficiency components such as reading comprehension (e.g., Farvardin and Koosha 2011; Qian 1999), others indicate that depth is superior (e.g., Mehrpour et al. 2010; Qian 2002). Using a series of simple linear regression analyses (one predictor, one criterion), Qian (2002) found that vocabulary size (Vocabulary Levels Test, Nation 1983), depth (adapted WAT version) and a test measuring synonym knowledge (Test of English as a Foreign Language [TOEFL] Vocabulary Item Measure) explained 54-59% total variance in L2 reading comprehension (TOEFL reading comprehension subtest). Any combination of two predictors was better than one alone, suggesting both vocabulary size and depth are important for L2 reading comprehension. Although all three predictors had a roughly equal relationship with reading comprehension, Qian argued that because the WAT taps richer parts of vocabulary knowledge than other measures (namely synonymy, polysemy, collocation), its use in assessment would produce the most positive teaching/learning washback effect.

The extent to which these predictive powers of vocabulary size and depth generalise to other language competence aspects is an open issue. Indeed, Schmitt notes "it is an interesting, but unexplored, question whether the two would equally predict other kinds of language use" (2014: 939). One such area is the use of metaphor and other figurative language.

To our knowledge, only one study, by Azuma (2005), has investigated L2 vocabulary knowledge and MC correlations. Azuma found that MC tests of ability to write literal/figurative, *An/the X is a(n) adjective Y* sentences (MC-XYT) and literal/figurative idiom/proverb understanding and production (MC-RT and MC-PT) correlated positively with both vocabulary size (Vocabulary Levels Test, versions developed by Schmitt 2000; Schmitt et al. 2001) and depth (a specially developed polysemy measure after a simplified WAT version was discarded as too challenging). MC-vocabulary correlations varied for three different groups of approximately 60 Japanese university age L2 English learners (*rho* = .28-.75) but showed generally positive relationships between these constructs. However, differences in these group's correlations and probable overlapping variance complicate interpretation of how, exactly, the receptive and productive MC measures related to vocabulary size and depth. Additionally, issues with MC item content and phrasing and the restricted scope of construct targeted mean Azuma's findings are probably limited to the context investigated rather than offering more generalisable patterns.

2.3. L2 MC and proficiency

Intuitively, a positive relationship between L2 MC knowledge components and general proficiency would be expected, given that metaphor pervades large parts of the language system (Low 1988) and is relevant to all parts of communicative competence (Littlemore and Low 2006a; 2006b). Where fluid, skill-based components of MC are concerned, these seem to be more independent of language proficiency (Johnson and Rosano 1993; Littlemore 2010).

Both elicited and naturalistic enquiries report positive connections between L2 MC and proficiency. (Given the nature of our measures, we focus on elicited MC, for studies using more naturalistic MC see Hoang and Boers 2018; Littlemore et al. 2014; Nacey 2013; and further discussion in O'Reilly and Marsden 2021). With 75 L1 Chinese learners of English, Zhao et al. (2014) found that their institution's cloze-item and comprehension question reading test correlated positively with Azuma's (2005) MC-RT (r_s = .44), but not her MC-XYT (r_s = .01). This latter finding was attributed to the productive items in the MC-XYT, although given that receptive and productive competences are generally linked, this finding is somewhat puzzling.

In a similar study, Aleshtar and Dowlatabadi (2014) administered NourMohamadi's (2010) English Conventional Metaphor Proficiency Test (ECMPT) and the 2001 Oxford Online Placement Test (hereafter OOPT) to 60 L1 Persian undergraduate L2 English learners. The ECMPT consisted of six (15-item) sections relating to six types of cross-linguistic metaphor variation reported by Kövecses

(2003). The results showed strong MC-proficiency correlations for 'low' and 'high' proficiency groups, r = .77, and .72 respectively.[2] However, high and low proficiency groups were formed according to whether participants were freshman or juniors, with problematic overlapping OOPT scores between the groups. Additionally, it was uncertain whether the test measured written or spoken-mode MC and ECMPT's reliability from NourMohamadi's (2010) study was reported, rather than with the authors' own participants.

Standardised tests such as International English Language Testing System (hereafter IELTS), TOEFL, and OOPT purport to measure one unified construct (L2 proficiency). However, such tests have different formats, question types, and item foci, and hence may differ in how they relate to MC skills and knowledge. Whereas IELTS taps broad reading/writing/speaking/listening 'proficiency' skills via (for example) multiple-choice, matching, plan/map/diagram labelling, oral examiner interview, and essay tasks, OOPT contains no speaking or extended writing component, and uses only multiple-choice and gap-fill questions to measure knowledge of grammatical forms, and semantic and pragmatic meanings encoded in social interactions.

Explicit acknowledgement of metaphor in these tests is minimal; for OOPT '…figurative meanings, metaphor' are only listed under 'Sociocultural meaning' but would apply to other 'Pragmatic meanings' sub-components (e.g., affective stance, humour) and indeed 'Grammatical meaning' (e.g., polysemy, collocation) and 'Grammatical form' (e.g., co-occurrence restrictions, word coinage) (Purpura 2004; 2021). In CEFR, metaphor/figurative language only begins to appear at C1 – Advanced (Proficient user – Effective Operational proficiency) where the learner "…is likely to demonstrate a [C2 = consistently] high level of communicative meaningfulness [and] be able to use metaphor, idioms, and colloquialisms, to convey finer shades of meaning" (Purpura 2021; see also Nacey 2013).

Metaphor/figurative language are even rarer in IELTS (O'Reilly 2017). Although linguistic metaphor logically sits within the 'lexical resource' scoring component (Writing and Speaking), IELTS scoring descriptors do not mention the relevance of metaphor/figurative language to L2 proficiency, even at higher levels, and these terms retrieve no search results on its main website. Although research exists on how and where IELTS taps pragmatic knowledge and (sub)competences (Allami and Aghajari 2014), we know of no studies on the relationship between L2 MC and IELTS.

Despite certain limitations, the empirical studies summarised in this section highlight that while MC and general proficiency are undoubtedly related, the strength of relationship may vary according to the exact measures used, specific MC constructs targeted, proficiency of learners, and confounding

[2] We interpret correlation strength using Cohen's (1988) criteria: .10-.29 (small); .30-.49 (medium); .50-1.0 (large).

effect of vocabulary knowledge. The latter (to our knowledge) has never been controlled for when estimate MC-proficiency test correlations.

2.4. Should L2 MC be taught and tested in practitioner contexts?

Metaphor is an important component of pedagogical toolkits, effective for reinvigorating stagnant learning environments (Macarthur 2010) and helping achieve syllabus aims. Above and beyond simply improving vocabulary size and depth, there could be value in helping learners use their current knowledge and contextual clues to work out and use metaphors that are (for them) novel, thus gaining a degree of autonomy over the learning process (Littlemore and Low 2006a; Littlemore 2002; 2004c). In this view, it would seem sensible for teachers to also collect data on (i.e., measure/test) learners' MC progress via elicited and naturalistic means (see Section 2.3).

Low/Littlemore argued in favour, generally, of the introduction of metaphor into the L2 classroom, drawing out pedagogical implications from their theoretically motivated MC descriptions. Low (1988) suggested tasks involving situational and/or linguistic constraints to teach creative metaphor(s), thus mitigating difficulties with too much creative freedom, and multi-texts/tasks involving different modes, registers, and styles for teaching conventional (or commonly used) metaphor. Littlemore and Low (2006a) frequently presented *implications for foreign language learning*, detailing classroom activities in which, for example, learners explore and argue via conceptual metaphors when debating (illocutionary competence), or practice using metaphor/sayings/proverbs to close a topic, smoothly and inoffensively (textual competence).

Pedagogical metaphor research has investigated the benefits of providing etymological elaboration, and highlighting alliteration, phonological properties, and conceptual relationships underpinning various expressions. This research shows mixed findings; in Boers et al. (2014) adult south-east Asian L2 English learners performed poorly, even decreasing in scores, through metaphor-laden verb-noun collocation practice exercises, while in Boers et al. (2007) adult L1 Dutch L2 English learners receiving etymological information improved their comprehension, recall, and sensitivity to figurative idiom register. Although teaching the conceptual underpinnings of metaphor has shown tentative results, this in itself does not offer a "magic bullet" (Macarthur 2010: 158) for helping learners develop MC, partly since not all linguistic metaphors are amenable to cognitive elaboration, and such approaches are not equally effective for all ages, proficiencies, and learning styles (Nacey 2013).

Moreover, teachers and learners may see metaphor as irrelevant, feel ill-equipped to deal with it, be frustrated by the lack of research-informed materials on 'system' rather than 'item' learning, overwhelmed and confused by what they find, and restricted in terms of time and freedom to experiment. TESOL practitioners must also carefully consider the target norms/English varieties presented to learners, where appropriate using a plurilithic/Englishing/any-metaphor-that-works approach, or where stricter lexical conventions apply (e.g., healthcare discourse) present specific target

linguistic metaphors, leveraging prototypes to encourage wider metaphoric thinking (Hall 2014; O'Reilly and Marsden 2021).

The pedagogical perspectives summarised above point to an exciting metaphor in ELT research agenda. However, we hasten to clarify that the analyses below are primarily concerned with further unravelling the nature of the MC construct and its relationship to other aspects of language competence, not with setting out an effective pedagogy, determining casualty, or prescribing how teachers might intervene in learning processes. Thus, we later provide *pedagogical reflections* rather than strong teaching implications.

2.5. Summary of research gaps, research questions

As we have argued, L2 learners have much to gain by becoming more proficient users of metaphor. However, questions about how MC develops alongside vocabulary size and depth, and how closely its various facets are connected to different measures of language proficiency remain unanswered. The evidence suggests an interrelatedness of MC knowledge components, and vocabulary knowledge and general proficiency measures, while more fluid, skill-based MC components may come into play at any stage of learning and do not seem correlate with proficiency at all (Section 2.3). However, substantive and methodological limitations to previous studies complicate interpretation and more robust approaches that isolate unique relationships by controlling for confounding interrelatedness are needed. This, in turn, will provide a more precise understanding of what learners can likely do with metaphor at different proficiencies, helpful for interpreting explicit and implicit references to metaphor in proficiency descriptors of large-scale tests. While vocabulary size and depth are relatively equally important parts of L2 proficiency, no study has investigated their respective roles in receptive or productive MC.

In the subsequent investigation, we address these research gaps using an observational/correlation design, taking a detailed snapshot of different levels of MC, vocabulary knowledge, and proficiency in a large group of homogenous learners (see Section 3.1). Two research questions were formulated:

RQ1: To what extent do various L2 MC measures overlap with two standardised L2 proficiency measures, controlling for L2 vocabulary knowledge?

RQ2: To what extent do L2 vocabulary size and depth relate, as statistical predictors, to various criterion L2 MC measures?

For both questions, 'various L2 MC measures' refers to the overall/composite receptive and productive measures and MC factors uncovered in O'Reilly and Marsden (2021). In the following section we provide details about the participants, measures used, and approaches taken to isolate the relationships of interest.

3. Method

All data collection and analysis materials created for this study are available in the Open Science Framework: https://osf.io/35czh/?view_only=03a4b4d642574d6eb107fe2e811ab101 and www.iris-database.org.

3.1. Participants

Participants were 108 L1 Mandarin speakers of L2 English (97 females), aged 18 to 31 years (M = 23.0, SD = 2.6), who completed the MC, vocabulary knowledge, and proficiency tests (see Section 3.2). Most (97/108) were UK university postgraduates at nine universities, engaged or enrolled in study; the remainder were undergraduates studying at these universities. All except seven were studying social science degrees. The tests were administered to a further four participants, later removed as extreme cases (scores < $M - 3 \times SD$ for Test 2-Metaphor Layering-R, V_YesNo, WAT, or IELTS). All participants reported learning English as a foreign language at school in China, from as young as 3 (starting age M = 9.2, SD = 2.6).[3] The average (median) reported time spent living in the UK was two months (interquartile range = 9) with most participants having arrived relatively recently (87% within 12 months).[4] Time spent in the UK was controlled for in all analyses reported below. A further 16 participants started the study but later dropped out due to other commitments.

3.2. Materials

MC tests/measures: The six MC measures used in the analyses below were obtained from O'Reilly and Marsden's (2021) MC Test Battery and exploratory factor analysis (see Section 1). Table 1 presents these six measures, the receptive (-R) or productive (-P) MC tests that contributed to them (and in the cases of factors, their strength of loading), what they were designed to elicit, and the number of items per test. All receptive tests (except two, see Table 1 notes) used the multiple-choice, gap-fill questions. All productive tests used limited production, gap-fill questions.

[3] The age of 18, reported by one participant, was omitted here. Qi (2016) notes that English was introduced in China as compulsory subject from Primary Three (i.e., age 8), when this participant was 10, suggesting they likely misreported this information.

[4] Four participants who had not spent any time in the UK comprised three commencing postgraduate studies and one commencing undergraduate studies within the next few weeks. Five participants had spent over two years in the UK and one of these over three years.

Table 1: Metaphoric competence (MC) measures from O'Reilly and Marsden (2021).

Measure	Contributing MC tests (loading)	Test of ability to	*k*
Receptive metaphoric competence		***Composite (observed) measure of all MC Test Battery receptive tests***	
	T1-Phrasal verbs-R	recognise metaphorical phrasal verb particles	10
	T2-Metaphor layering-R[a]	understand metaphors, recognise meanings & garden path endings	10
	T3-Vehicle acceptability-R[b]	rate the acceptability of semantic and word class exploitations of Vehicles	18
	T4-Topic/Vehicle-R[b]	rate the acceptability of Vehicles as analogies for given Topics	8
	T5-Topic transition-R	recognise idioms/proverbs/sayings in topic transition	8
	T6-Heuristic-R	recognise similes used to perform heuristic functions	9
	T7-Feelings-R	recognise metaphors that convey feelings about information	8
	T8-Idiom extension-R	recognise extensions of the literal senses of idioms	11
	T9-Metaphor continuation-R	recognise continuations of metaphor in discourse	10
Productive metaphoric competence		***Composite (observed) measure of all MC Test Battery productive tests***	
	T1-Phrasal verbs-P	recall metaphorical phrasal verb particles	10
	T5-Topic transition-P	produce idioms/proverbs/sayings in topic transition	8
	T6-Heuristic-P	produce similes to perform heuristic functions	8
	T7-Feelings-P	produce metaphors that convey feelings about information	8
	T8-Idiom extension-P	produce extensions of the literal senses of idioms	12
	T9-Metaphor continuation-P	produce continuations of metaphor in discourse	9
Productive Illocutionary MC		***Factor 1 (latent), explaining 13% total variance in 15 MC tests***	
	T6-Heuristic-P (.64)	produce similes to perform heuristic functions	8
	T7-Feelings-P (.60)	produce metaphors that convey feelings about information	8
	T2-Metaphor layering-R (.50)	understand metaphors, recognise meanings & garden path endings	10
	T7-Feelings-R (.37)	recognise metaphors that convey feelings about information	8
	T5-Topic transition-P (.37)	produce idioms/proverbs/sayings in topic transition	8
Metaphor Language Play		***Factor 2 (latent), explaining 10% total variance in 15 MC tests***	
	T8-Idiom extension-P (.90)	produce extensions of the literal senses of idioms	12
	T8-Idiom extension-R (.41)	recognise extensions of the literal senses of idioms	11
	T9-Metaphor continuation-P (.38)	produce continuations of metaphor in discourse	9

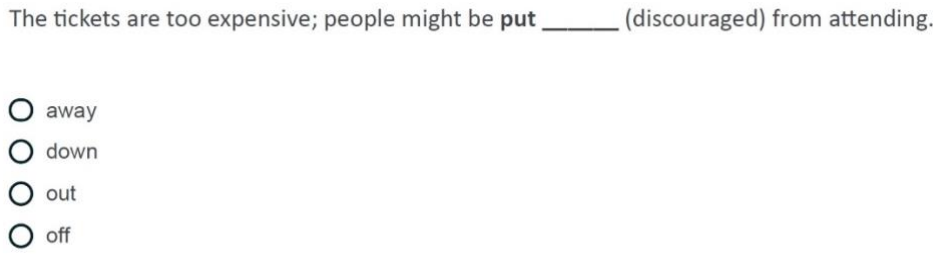
Topic/Vehicle Acceptability	***Factor 3 (latent), explaining 7% total variance in 15 MC tests***	
T3-Vehicle acceptability-R (.66)	rate the acceptability of semantic and word class exploitations of Vehicles	18
T5-Topic transition-R (.33)	recognise idioms/proverbs/sayings in topic transition	8
Grammatical MC	***Factor 4 (latent), explaining 5% total variance in 15 MC tests***	
T1-Phrasal verbs-P (.59)	recall metaphorical phrasal verb particles	10
T1-Phrasal verbs-R (.43)	recognise metaphorical phrasal verb particles	10

Notes. 108 participants were retained in the current study. Abbreviations: MC = metaphoric competence; -R = receptive test; -P = productive test.

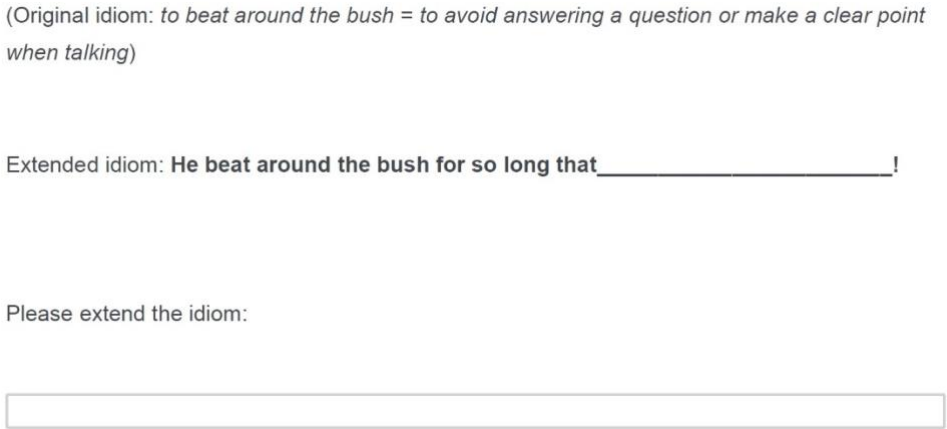
[a] Also had limited production (explain-the-meaning) questions.
[b] Rating scale (acceptability judgement) questions.

The first two MC measures were participants' composite scores from all receptive and all productive tests in MC Test Battery. The remaining four MC measures comprised participants' factor scores, calculated using Thurstone's (1935) regression method since three out of four factors were significantly, positively correlated. Examples of MC test items include:

Test 1-Phrasal verbs-Receptive, multiple-choice (correct answer = 'off')



Test 8-Idiom extension-Productive (open gap-fill)



(See further selected examples in O'Reilly and Marsden 2021, and full instrument and scoring procedures in www.iris-databse.org).

Instrument reliability analyses suggested high internal consistency of items for Overall Receptive MC, Overall Productive MC, Productive Illocutionary MC, Metaphor Language Play, and

Topic Vehicle Acceptability (ordinal omega *M* = .81, *SD* = .05), somewhat lower for Grammatical MC (.6 range).[5] Agreement between raters scoring the productive responses was *substantial* (weighted kappa = .62 and .66 for two sets of comparisons), and intrerrater reliability was *almost perfect* (weighted kappa = .81) by Landis and Koch's (1977) criteria.

Vocabulary size: Our conceptualisation of vocabulary size was participants' ability to correctly identify written, decontextualized real and imaginary words, as measured by Meara and Miralpeix's (2015) widely used V_YesNo test. The V_YesNo (administered via www.lognostics.co.uk) requires test-takers to respond to 200 words, clicking 'yes' if a word is known and 'no' if not or if unsure. A single vocabulary size estimate is provided at the end. Behind the scenes, the system analyses ten blocks of 20 words, each with 10 real and 10 pseudo words that serve to correct for overestimated knowledge.

Compared with other multiple-choice vocabulary size measures, the V_YesNo is easily scored (Pellicer-Sánchez and Schmitt 2012), maximises word coverage while minimising reading burden, and better discriminates between learners with higher and lower vocabulary sizes (Meara and Buxton 1987). With the Vocabulary Levels Test (see Section 2.2) and multiple-choice tests generally, a learner who knows the target word may answer incorrectly due to unfamiliarity with words contained within the context or definitions, or the particular meaning targeted despite other meanings being known (Meara and Buxton 1987). With such tests, the number of items needs to increase for learners with higher vocabulary sizes to keep the proportion of known vocabulary tested (test coverage) constant. The V_YesNo avoids both of these problems. Both pilot and main study participants reported appreciating the test's efficiency, automated/immediate scoring, and professional interface (O'Reilly 2017), suggesting good face validity (Read 2000).

Instrument reliability, measured as the internal consistency of continuous scores for each of the ten blocks (see above) was high (.9) by Omega (total), Revelle's Omega (total), Greatest Lower Bound, and Coefficient *H*, four superior alternatives to Cronbach's alpha (McNeish 2018).[6] Additionally, all 108 participants scored higher than 2,500, the point below which estimates may become unreliable (Meara and Miralpeix 2015).

Vocabulary depth: Our conceptualisation of vocabulary depth was lexical organisation, specifically, learners' ability to recognise written, decontextualised semantic, and collocation word associations as measured by Read's (1998) WAT (Section 2.2, available in www.iris-databse.org). The WAT offers a

[5] All reliability estimates were calculated using the optimised MC Test Battery items (versions 1 and 2) presented in O'Reilly and Marsden (2021) and the 108 participants retained in this study. Grammatical MC's comparatively lower internal consistency may be because of its smaller number of items. Even so, results concerning this measure should be interpreted cautiously.

[6] At the time of data collection, only overall scores were recorded. We express our sincere gratitude to the website maintainer for subsequently retrieving the full item data for 105/112 test-takers and are fully responsible for the irretrievable cases. Since four test-takers were deleted from the study as extreme cases (see Participants), and a further six cases were irretrievable, V_YesNo reliability estimates are reported from 102 cases.

detailed picture of learners' lexical networks and, as the most commonly used (but by no means only) vocabulary depth format (Schmitt 2014), allowed us to better compare findings with other key studies (e.g., Qian 2002). Given the testing burden required to obtain robust, nuanced measures of MC (the less well-attested construct), we opted not to operationalise vocabulary size and depth in a multitude of different ways (e.g., as in Webb 2005), but certainly advocate this exploration in future research.

For each item, an adjective headword is presented with eight possible associates, four adjectives and four nouns. Test-takers must identify adjectives with paradigmatic or analytic associations with the headword or nouns with syntagmatic associations (Read 1993; 1998). The test taps knowledge of the main meanings of polysemous words. Variation in the location of correct associates among the distractors in the 1998 WAT make it less susceptible to random guessing than previous iterations (Schmitt et al. 2011). It also has advantages over the Vocabulary Knowledge Scale (Wesche and Paribakht 1996), which requires more reading and risks eliciting ambiguous productions. The WAT was developed as a reliable measure (Read 1993; 1998) and had high internal consistency with our participants (.8-.9) by Omega (total), Revelle's Omega (total), Greatest Lower Bound, and Coefficient *H* assuming a continuous scale (McNeish 2018).

L2 proficiency: We used the OOPT and IELTS as L2 proficiency measures. These two tests differ in item format and focus, and how they operationalise proficiency (see below and Section 2.3). It was advantageous to include *both* these high stakes tests as correlates since this allowed for investigation into their unique relationships with MC, it provided more exhaustive proficiency coverage, allowed for a wider comparison with existing research, increased the relevance of the study to more learning contexts, and achieved a better balance between the number of proficiency and vocabulary knowledge measures investigated.

The OOPT was designed to allow for a quick and reliable measure of a learners' general language ability and proficiency level placement. Via multiple-choice and limited-production gap-fill, the test's Use of English and Listening sections contain questions that measure knowledge of grammatical forms, semantic meaning, and pragmatic meanings encoded in social interactions (e.g., implied meanings). The OOPT is computer-adaptive, selecting questions from a large bank of standardised items that have been extensively piloted for reliability. A correct answer results in a more difficult following question, an incorrect answer an easier one. While technically a placement test, the OOPT's robustness and theoretical underpinning (Pollitt 2021; Purpura 2021) made it suitable for present purposes.

By comparison, 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 skills via various question types (Section 2.3) and purports to "actively [avoid] cultural bias, and [accept] all standard varieties of native speaker English, including North American, British,

Australian and New Zealand English" (www.ielts.org). Given that all our participants had completed IELTS as a condition of studying in the UK, and the impracticality of having them repeat it, we recorded scores from their most recent attempt. Institutions generally treat IELTS scores as valid for two years. In our sample, IELTS test dates were provided by 100/108 participants, the majority of participants (79%) had taken IELTS within the last two years (median months = 14, interquartile range = 13).[7] To optimise the validity of the IELTS measure and maximise statistical power, we controlled for the heterogeneity in time lapse between IELTS and present study scores in the various analyses rather than (for example) imposing two years as an arbitrary cut-off for participant retention. This decision was also informed by research showing modest or negligible English development in UK-based L1 Mandarin speakers over the course of an academic year, and that IELTS remains predictive of academic outcomes even if taken at different points prior to arrival and regardless of whether learners had completed a three-month pre-sessional English programme prior to their main programme (Trenkic and Warmington 2019).

3.3. Procedure

Main data collection took place from June to November 2015. All participants completed the MC Test Battery online (via Qualtrics), receiving £5 cash or equivalent value Amazon voucher. For logistical reasons and to minimise test-taker anxiety, observed in previous L2 MC studies, participants were given the choice of completing the MC Test Battery, vocabulary tests and OOPT in lab sessions, or at home in their own time. Approximately one third (n = 35) attended lab sessions, and two thirds (n = 73) completed tests at home, with a subsequent analysis confirming no detectable score difference between settings.[8] In the lab, participants completed the MC Test Battery in 1.5 to 2 hours (similar to Littlemore 2001) took a scheduled 15-minute break, and proceeded the V_YesNo, WAT and OOPT, taking a further 1.5 hours approximately. To mitigate fatigue, participants were encouraged to take short comfort breaks and enjoy refreshments throughout the testing session. Participants provided IELTS scores and were informed that the tests were unconnected to their studies and that they could withdraw at any point.

3.4. Data analysis

Data were analysed using R programming language (R Core Team 2016) with 20 packages/scripts and Microsoft Excel. To address RQ1, we ran a total of 12 partial correlation analyses to estimate the specific relationships between the six MC measures shown in Table 1 (y-axis) and two proficiency measures (x-axis) whilst controlling for variance in these bivariate relationships explained by the other proficiency measure (the one not in focus), vocabulary size and depth, months spent in the UK and

[7] Of the remaining 21 participants, 18 had taken IELTS within the last two and a half years, a further two within three years, and one 48 months prior.

[8] Munzel and Brunner's (2000) method of robust MANOVA showed no significant main effect of test setting on MC test scores, $F(2,15) = 0.73$, $p = .685$, with fairly similar score ranks in all cases.

months since IELTS.[9] For the most part, data met parametric assumptions and so Pearson's *r* was used; where a main (i.e., non-control) correlate did not meet parametric assumptions, Spearman's r_s was used. Bonferroni adjustment for 12 repeated tests (i.e., dividing .05, .01, and .001 by 12) was employed to help interpret the significance of estimates.[10] For both proficiency measures, we used the overall score rather than components (e.g., OOPT Use of English, IELTS reading) for three main reasons: (1) we wanted to understand how proficiency, as a sum-of-all-skills, relates to the MC construct; (2) with 108 participants, six (rather than two) proficiency variables would mean an increase from 12 to 36 correlations, requiring highly strict alpha value adjustment for repeated testing and reduced power to detect relationships (Miles and Shelvin 2001); (3) introducing more IELTS-based variables did not seem sensible, given the need to statistically control for possible confounds related to these data.

To address RQ2 we employed multiple regression to estimate the extent to which vocabulary size and depth statistically predict, in combination and individually, scores in the MC criterion measures. We used multiple regression, rather than partial correlation, because the vocabulary knowledge literature provides a good motivation for extending the exploration into how vocabulary size and depth predict different types of language use (Qian 2002; Schmitt 2014); because this enabled us to estimate the amount of common and unique MC variance explained by these two vocabulary measures; and because the MC Test Battery was *more* than just a vocabulary size or depth test, tapping whole-sentence comprehension and production (i.e., sentential comprehension, grammatical knowledge etc.) and multiple semantic relations in one test item rather than knowledge connected to isolated words. Thus, the larger, richer, more complex (and less well-attested) set of competences were modelled as a function of the more conceptually parsimonious, controlled (and better-attested) vocabulary constructs. Although multiple regression requires an explicit decision about which variables should be entered as predictors and criterion, it estimates *relationships*, not causality (Winter 2020), which was not inferable with our observational/correlational research design.

A total of six multiple regression analyses were run, each using the same three predictors (V_YesNo, WAT, and as a control measure, months spent in the UK), but with a different MC criterion measure from the six in Table 1. Bonferroni adjustment for six repeated tests helped interpret the significance of estimates. Since V_YesNo and WAT use different metrics, we first standardised scores for these two variables by centering the means to zero and dividing each score by the standard deviation, allowing us to directly compare their regression coefficients.

[9] Interested readers can obtain vocabulary knowledge-proficiency correlations using the Open Science Framework materials (start of Section 3).

[10] Although Bonferroni and other adjustments reduce Type I error rate, their usefulness has been questioned (e.g., Perneger 1998). While statistical significance (after Bonferroni adjustment) helped highlight our most salient findings, we also report and consider 95% confidence intervals when interpreting estimates.

4. Results

4.1. Descriptive statistics

Table 2 presents measures of average and spread for the various tests, the number of items, and the scales used.

Table 2: Descriptive statistics and variables overview.

Test/Variable (regression)	*K* items	Scale	Mean	SD
Vocabulary knowledge and proficiency measures				
V_YesNo Test	200	0 - 10,000 (words)	5916.28	1187.54
WAT	160[a]	0 - 160 (associates)	126.37	10.16
OOPT	Varies[b]	0 - 120 (points)	67.06	13.96
IELTS	Varies[c]	0 - 9 (bands)	6.64	0.52
Months since IELTS (control)	_	months	14[d]	13[e]
Months in UK (control)	_	months	2[d]	9[e]
Metaphoric competences (MC) measures				
Overall Receptive MC[f]	9[g]	0-100(%)	54.54	12.16
Overall Productive MC[f]	6[g]	0-100(%)	43.84	17.69
Metaphor Language Play	15[g]	Z-scores (approx. -3.0 to +3.0)	0.0019	0.8614
Productive Illocutionary MC	15[g]	Z-scores (approx. -3.0 to +3.0)	0.0259	0.9179
Topic/Vehicle Acceptability	15[g]	Z-scores (approx. -3.0 to +3.0)	-0.0008	0.7800
Grammatical MC	15[g]	Z-scores (approx. -3.0 to +3.0)	-0.0006	0.7412

[a] 40 headwords each with four correct associates.
[b] Computer adaptive.
[c] Ten Listening and 40 Reading questions, two Writing tasks, various Speaking questions.
[d] Median.
[e] Interquartile range.
[f] Values differ slightly to O'Reilly and Marsden (2021), who retained data from 109 participants (receptive) and 112 participants (productive).
[g] Number of metaphoric competence tests submitted to factor analysis.

Figure 1 (below) shows the concentration of scores for the MC measures. The large circles show means, and the bars indicate the range of one standard deviation above and below this.

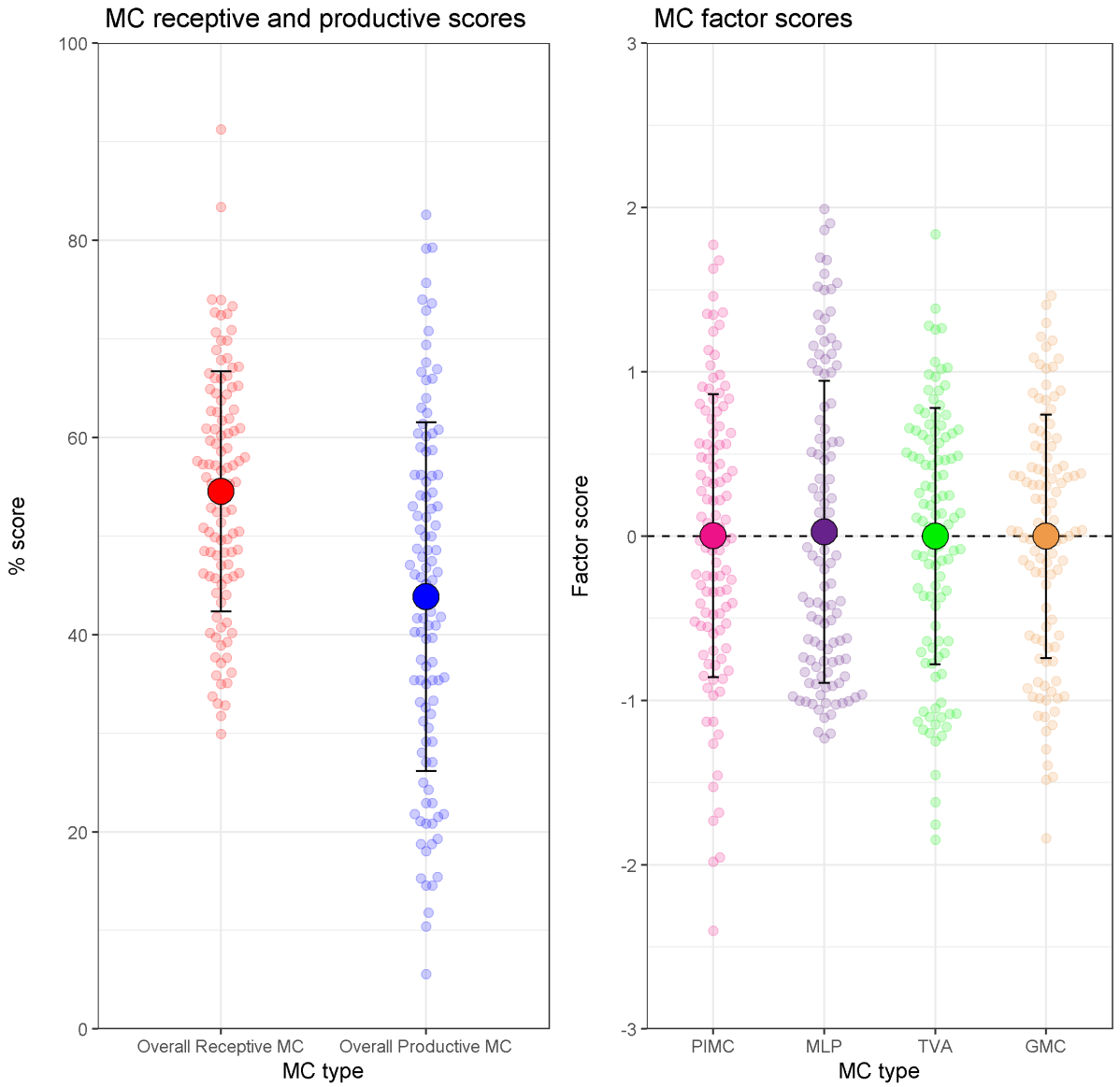


Figure 1: Overall Receptive and Productive MC scores (left) and MC factor scores (right): PIMC = Productive Illocutionary MC, MLP = Metaphor Language Play, TVA = Topic/Vehicle Acceptability; GMC = Grammatical MC; small dots = individual scores, larger points = group mean scores, vertical black bars = +/- 1 × standard deviation, *N* = 108.

Table 2 and Figure 1 show that, on average, participants had a vocabulary size of just under 6,000 words, a vocabulary depth extending to approximately 80% associates recognised, were at the lower B2 CEFR level according to OOPT, and were between IELTS 6.5 and 7.0. Overall Receptive MC scores were higher than Overall Productive MC, while MC factor scores show the most variation for Productive Illocutionary MC and the least for Grammatical MC.

4.2. Correlation analyses (RQ1): MC and proficiency

Bivariate correlations (without controls): Figures 2 and 3 present scatterplots and estimates (above each plot) of the simple correlations between Overall Receptive and Productive MC and proficiency

measures (Figure 2), and the four MC factors and proficiency measures (Figure 3) without controls. The black dots show participants' scores, the black trend lines a simple linear fit assuming a constant rate of increase across the data for each analysis, while the red lines (and the shaded areas, denoting the standard error) show a localised, polynomial fit where the rate of increase varies in different parts of the data, particularly the outer edges.

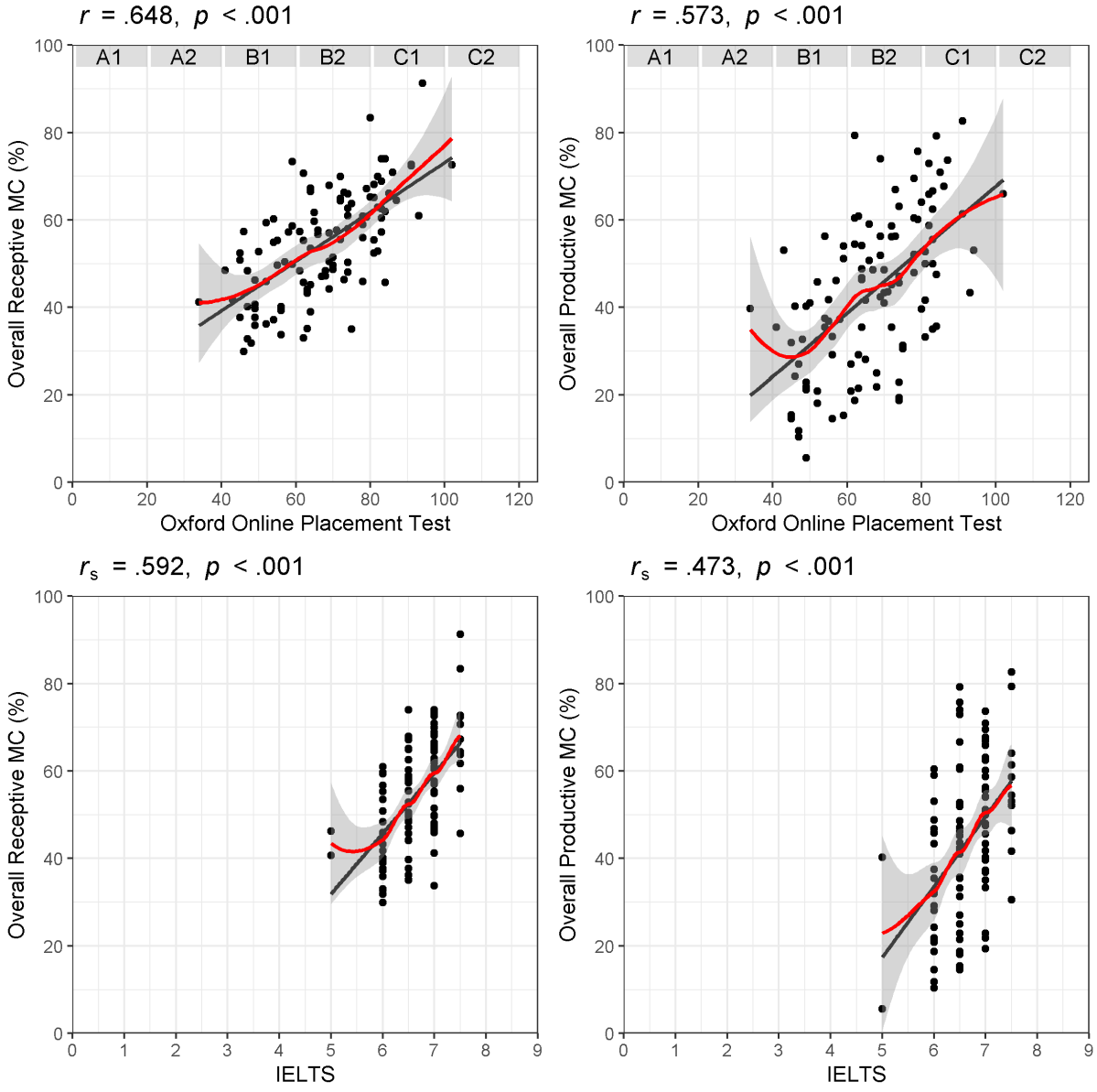


Figure 2: Scatterplots showing bivariate correlations (without controls) between Overall MC measures (*y*-axis) and general proficiency measures (*x*-axis); black dots = participant scores; black line = trend with linear fit; red line = trend with 'local' polynomial fit; A1, A2 etc. = CEFR proficiency level; *N* = 108.

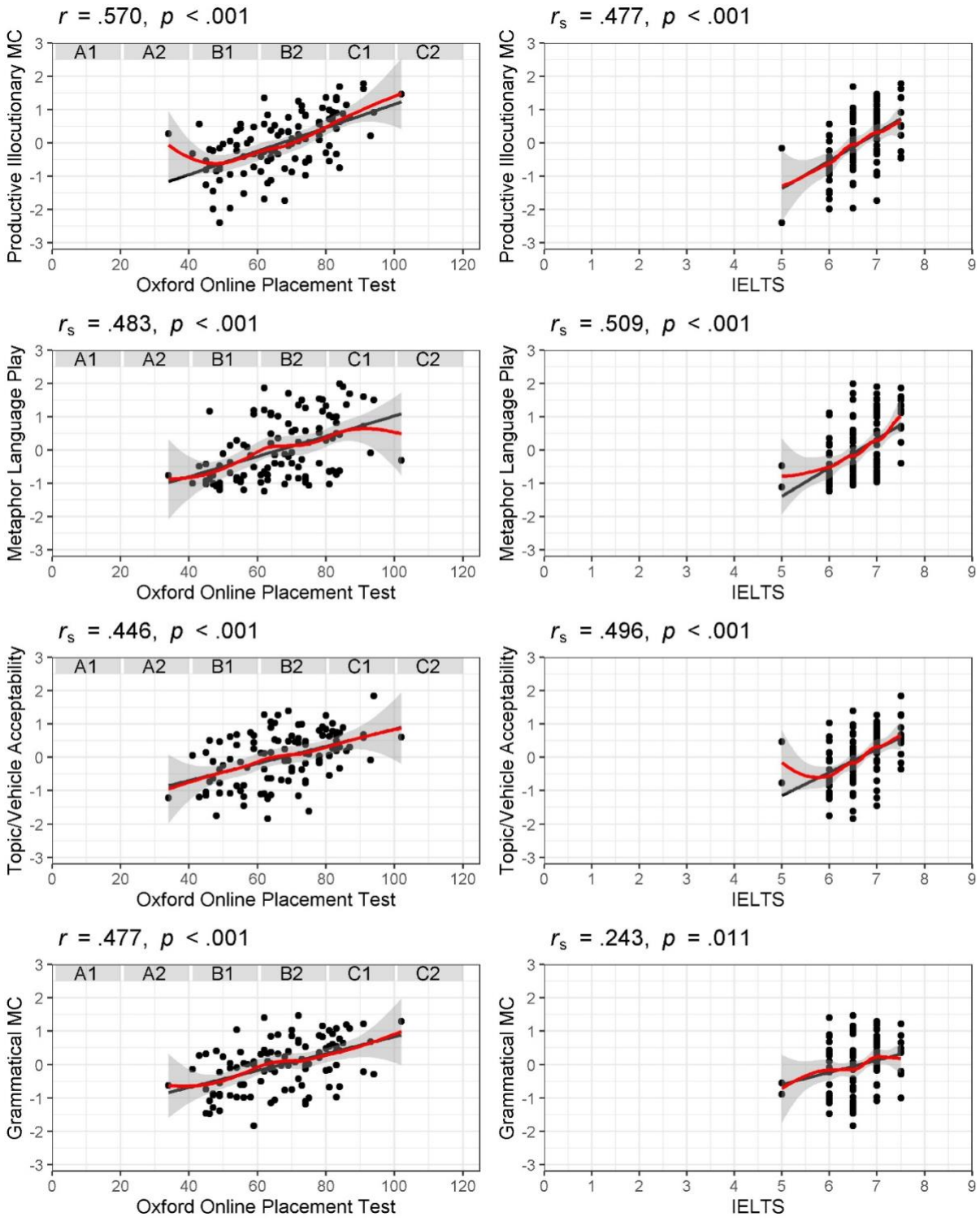


Figure 3: Scatterplots showing bivariate correlations (without controls) between MC factor scores (*y*-axis) and general proficiency measures (*x*-axis), for figure details see Figure 2.

The results showed that all MC and proficiency variables were positively and significantly correlated. Correlations between the Overall MC and general proficiency measures (Figure 2) were all large, with the exception of Overall Productive MC and IELTS, which was medium. Overall Receptive MC had a

stronger relationship with both proficiency measures than Overall Productive MC, while the OOPT had a stronger relationship with both Overall MC measures than IELTS did. The MC factors also all had positive, statistically significant correlations with the proficiency measures (Figure 3). These correlations were mostly in the medium to large range, while the relationship between Grammatical MC and IELTS was small. Unlike the Overall MC measures (Figure 2), the MC factors differed in their most closely associated proficiency measure; Productive Illocutionary MC and Grammatical MC were more strongly linked to OOPT, and Metaphor Language Play and Topic/Vehicle Acceptability to IELTS (Figure 3). In passing, we note that Overall Receptive MC and Overall Productive MC had a large positive correlation (r = .64, p < .001, not shown above), a stronger relationship between receptive and productive MC than in Azuma's (2005) study (r_s = .33, .53, .37, n = 57, 56, 59).

Partial correlations: Table 3 presents the original bivariate correlation estimates followed by partial correlation statistics including estimates (default r, for nonnormally distributed main variables r_s), their 95% confidence intervals (CIs), squared estimates showing percentage of shared variance (default R^2, for nonnormally distributed variables R_s^2, i.e., shared variance of ranked scores), the control measures list, and the p-value with asterisks showing statistical significance in line with Bonferroni adjustment for the 12 repeated tests.

Table 3: Partial correlations (Pearson's r default, Spearman's r_s for nonnormally distributed variables[a]) showing unique relationships between MC and proficiency measures with controls[b] (N = 108)

			Partial correlation				
x-axis	y-axis	Bivariate correlation estimate	Estimate	95% CIs for estimate	Estimate^2 (%)	Controls type[b]	p
OOPT	Overall Receptive MC	.648	.378	[.199,.532]	.143 (14.3%)	1	<.001**
IELTS	Overall Receptive MC	.592[a]	.377[a]	[.198,.532]	.142 (14.2%)	2	<.001**
OOPT	Overall Productive MC	.573	.326	[.142,.489]	.107 (10.7%)	1	<.001**
IELTS	Overall Productive MC	.473[a]	.199[a]	[.006,.378]	.040 (4%)	2	.044
OOPT	Productive Illocutionary MC	.570	.293	[.106,.461]	.086 (8.6%)	1	.003*
IELTS	Productive Illocutionary MC	.477[a]	.209[a]	[.017,.387]	.044 (4.4%)	2	.034
OOPT	Metaphor Language Play	.483[a]	.192[a]	[-.002,.372]	.037 (3.7%)	1	.052
IELTS	Metaphor Language Play	.509[a]	.275[a]	[.086,.445]	.076 (7.6%)	2	.005
OOPT	Topic/Vehicle Acceptability	.446[a]	.134[a]	[-.061,.319]	.018 (1.8%)	1	.177
IELTS	Topic/Vehicle Acceptability	.496[a]	.296[a]	[.108,.463]	.087 (8.7%)	2	.002*
OOPT	Grammatical MC	.477	.328	[.144,.490]	.108 (10.8%)	1	<.001**
IELTS	Grammatical MC	.243[a]	.037[a]	[-.157,.229]	.001 (0.1%)	2	.709

[a] Spearman's r_s used because IELTS, Metaphor Language Play and Topic/Vehicle Acceptability nonnormally distributed.
[b] Controls type 1 = V_YesNo, WAT, IELTS, UK months, IELTS months; controls type 2 = V_YesNo, WAT, OOPT, UK months, IELTS months.
*p < .0042, **p < .0008, ***p < .0001, Bonferroni adjustment to .05, .01, and .001 alpha levels applied for use of same controls in 12 correlations.

The partial correlation estimates are invariably smaller than the bivariate estimates, meaning the control measures explained part of the original bivariate relationships. Generally, with the various controls, Overall MC had higher amounts of shared variance with proficiency than the MC factors did. Overall Receptive MC shared 14.3% and 14.2% variance with OOPT and IELTS respectively, whereas Overall Productive MC shared 10.7% variance with the OOPT, but had a comparatively small overlap with IELTS, non-significant by the stricter Bonferroni adjusted level. Turning to the MC factors, only Productive Illocutionary MC and Grammatical MC had significant amounts of shared variance with the OOPT (8.6% and 10.8% respectively), while only Topic/Vehicle Acceptability had a significant overlap with IELTS (8.7% shared variance).

In sum, when controlling for the effects of vocabulary knowledge and other measures, the various aspects of MC did not universally and equally correlate with the proficiency measures. While certain types of MC appear to be more closely linked to certain types of proficiency (e.g., Grammatical MC and OOPT), other types (e.g., Metaphor Language Play) do not share a strong unique relationship with any general proficiency measure.

4.3. Correlation and multiple regression analyses (RQ2): MC and vocabulary knowledge

Bivariate correlations (without controls): Figures 4 and 5 present the simple correlations between the Overall Receptive and Productive MC and vocabulary knowledge measures (Figure 4) and between the four MC factors and vocabulary knowledge measures (Figure 5) without controls.

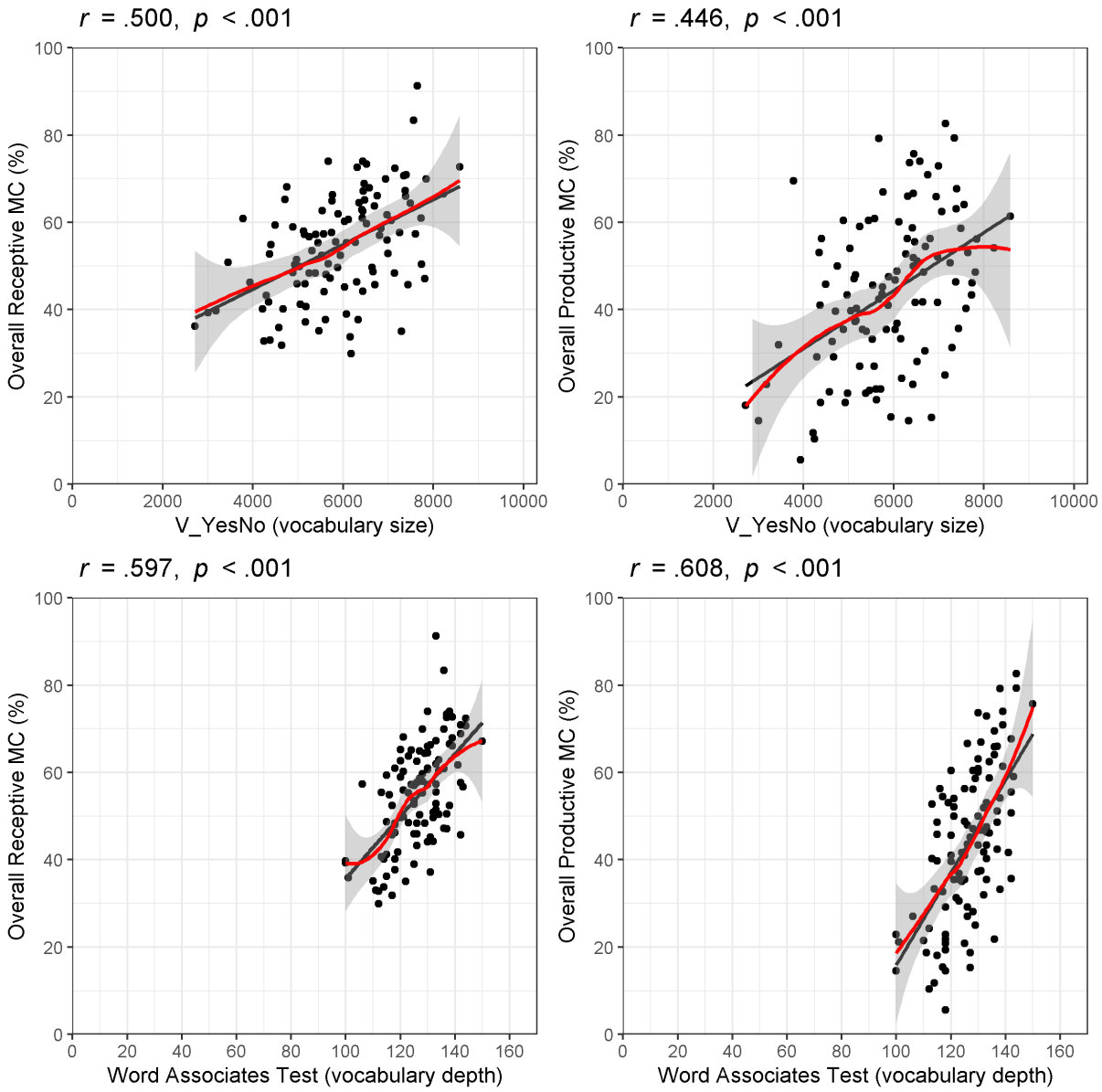


Figure 4: Scatterplots showing bivariate correlations (without controls) between Overall MC measures (*y*-axis) and vocabulary knowledge measures (*x*-axis), for figure details see Figure 2.

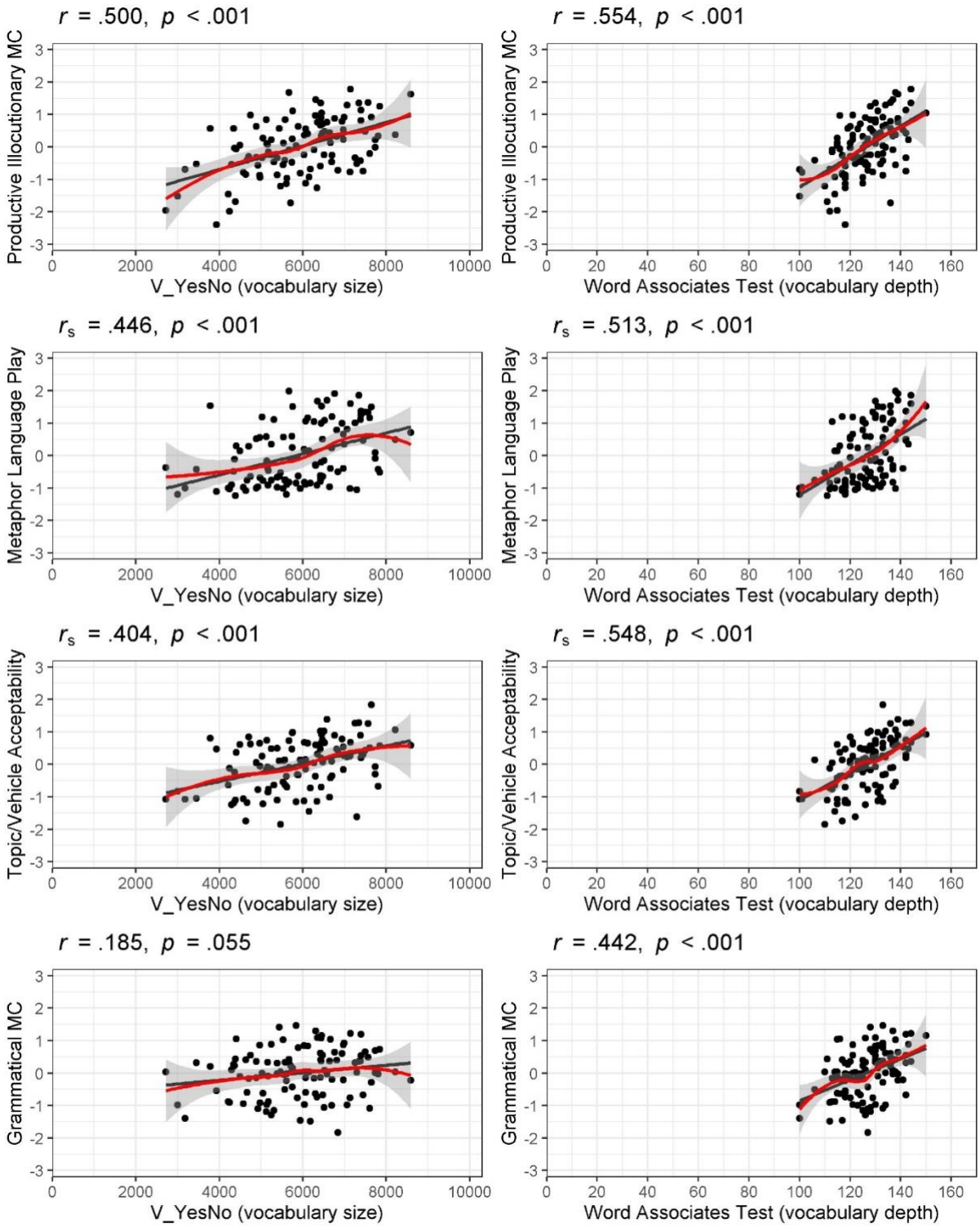


Figure 5: Scatterplots showing bivariate correlations (without controls) between MC factor scores (*y*-axis) and vocabulary knowledge measures (*x*-axis); for figure details see Figure 2.

The results showed that all MC measures were positively and significantly correlated with the vocabulary knowledge measures (except Grammatical MC and the V_YesNo), having large (in one case, medium) correlations. The WAT had a stronger association with both Overall MC measures than

V_YesNo and was particularly strongly related to Overall Productive MC. Overall Receptive MC was more closely related to V_YesNo than WAT. Correlations between the MC factors and vocabulary knowledge measures were medium-to-large (except Grammatical MC and V_YesNo). MC factors were invariably more strongly related to WAT than V_YesNo.

Multiple regression: In order to investigate the power of the combined and individual vocabulary knowledge measures to statistically predict variance in the MC measures we conducted six multiple linear regression analyses. In model 1 we regressed Overall Receptive MC (the criterion) simultaneously onto the standardised V_YesNo and WAT (the two predictors of interest), as well as months in the UK (for control purposes only). In model 2, we regressed Overall Productive MC onto these same predictors, while in models 3-6 we regressed the MC factors onto these same predictors.

In all cases, the models were significant compared with intercept only models, indicating statistically reliable effects of the predictors in combination. The adjusted (for the number of predictors) R^2 estimates, show that the three predictors accounted for 38% and 37% of the variance in Overall Receptive and Productive MC respectively, and 36%, 29%, 29%, and 20% in Productive Illocutionary MC, Metaphor Language Play, Topic/Vehicle Acceptability, and Grammatical MC respectively. Table 4 reports the standardised coefficients, their 95% confidence intervals (CIs), and beta values, which all provide information about each predictor's unique variance with the MC criterion with the effects of the other predictors partialled out, and the statistical significance of estimates with Bonferroni adjustment.

Table 4: Multiple regression analysis: Overall Receptive MC (models 1, 2) and MC factors (models 3-6) predicted by vocabulary knowledge and control measures, $N = 108$.

Model/Criterion	Predictors	Standardised estimate	95% CIs for estimate	Beta	*SE*	*t*	*p*
Model 1: MCR predicted	(Intercept)	54.10	[51.89,56.30]	_	1.11	48.59	<.001***
$R^2 = 0.40$, Adj. $R^2 = 0.38$	V_YesNo	2.94	[0.75,5.14]	0.24	1.11	2.66	.010
$F(4,104) = 23.28$	WAT	5.51	[3.30,7.71]	0.45	1.11	4.95	<.001***
$p < .001$	UK months	0.07	[-0.12,0.25]	0.05	0.09	0.70	.490
Model 2: MCP predicted	(Intercept)	43.24	[40.00,46.49]	_	1.64	26.45	<.001***
$R^2 = 0.39$, Adj. $R^2 = 0.37$	V_YesNo	2.76	[-0.47,5.98]	0.16	1.63	1.70	.090
$F(4,104) = 22.18$	WAT	9.06	[5.82,12.30]	0.51	1.63	5.54	<.001***
$p < .001$	UK months	0.09	[-0.19,0.36]	0.05	0.14	0.64	.520
Model 3: PIMC predicted	(Intercept)	-0.06	[-0.22,0.10]	_	0.08	-0.78	.440
$R^2 = 0.37$, Adj. $R^2 = 0.36$	V_YesNo	0.23	[0.07,0.39]	0.27	0.08	2.89	<.001*
$F(4,104) = 20.74$	WAT	0.33	[0.17,0.49]	0.38	0.08	4.10	<.001***
$p < .001$	UK months	0.01	[0.00,0.02]	0.11	0.01	1.41	.160
Model 4: MLP predicted	(Intercept)	0.11	[-0.07,0.28]	_	0.09	1.17	.240
$R^2 = 0.31$, Adj. $R^2 = 0.29$	V_YesNo	0.19	[0.01,0.37]	0.20	0.09	2.09	.040
$F(4,104) = 15.58$	WAT	0.40	[0.22,0.58]	0.43	0.09	4.40	<.001***
$p < .001$	UK months	-0.01	[-0.03,0.00]	-0.13	0.01	-1.57	.120
Model 5: TVA predicted	(Intercept)	0.03	[-0.12,0.18]	_	0.08	0.37	.710
$R^2 = 0.31$, Adj. $R^2 = 0.29$	V_YesNo	0.14	[-0.01,0.29]	0.18	0.08	1.85	.070
$F(4,104) = 15.85$	WAT	0.35	[0.20,0.50]	0.45	0.08	4.60	<.001***
$p < .001$	UK months	0.00	[-0.02,0.01]	-0.06	0.01	-0.68	.500
Model 6: GMC predicted	(Intercept)	-0.08	[-0.23,0.07]	_	0.08	-1.05	.300
$R^2 = 0.23$, Adj. $R^2 = 0.20$	V_YesNo	-0.07	[-0.22,0.08]	-0.10	0.08	-0.95	.350
$F(4,104) = 10.09$	WAT	0.34	[0.19,0.50]	0.46	0.08	4.45	<.001***
$p < .001$	UK months	0.01	[0.00,0.02]	0.16	0.01	1.83	.070

Notes. Abbreviations: Adj. = adjusted; MCR = Overall Receptive MC; MCP = Overall Productive MC; PIMC = Productive Illocutionary MC; MLP = Metaphor Language Play; TVA = Topic/Vehicle Acceptability; GMC = Grammatical MC.
*$p < .00833$, **$p < .00167$, ***$p < .00017$, Bonferroni adjustment applied to .05, .01, and .001 alpha levels for use of same predictors in six models.

Table 4 results show that increased WAT scores predict increased scores for all Overall MC and MC factor variables. Specifically, a one standard deviation increase in a participant's WAT score (roughly 10/160 more associates recognised, or 6.3%), holding the other predictors constant, would be linked to a 5.51% improvement in Overall Receptive MC and a 9.06% improvement in Overall productive MC, the equivalent of a participant increasing their Overall Receptive MC group ranking (out of 108) by about 9 places, and their Overall Productive MC group ranking (out of 108) by about 12 places. For the MC factors, the same WAT change suggests factor score increases of 0.33 (Productive Illocutionary MC), 0.40 (Metaphor Language Play), 0.35 (Topic/Vehicle Acceptability), and 0.34 (Grammatical MC), akin to group ranking increases of about 8, 13, 10, and 11 places (out of 108) for these factors respectively. V_YesNo shared unique variance with Productive Illocutionary MC only when interpreted using the stricter (adjusted) significance thresholds, although the 95% confidence intervals indicate it may have had some small, unique overlap with Overall Receptive MC and Metaphor Language Play. A one standard deviation increase in a participant's V_YesNo score then (roughly 1188 or 11.9% more V_YesNo words recognised), holding the other predictors constant, predicts a

Productive Illocutionary MC factor score increase of 0.23, a group rank improvement of about 7 places (out of 108) for this type of MC. These findings are visualised in Figures 6 and 7.

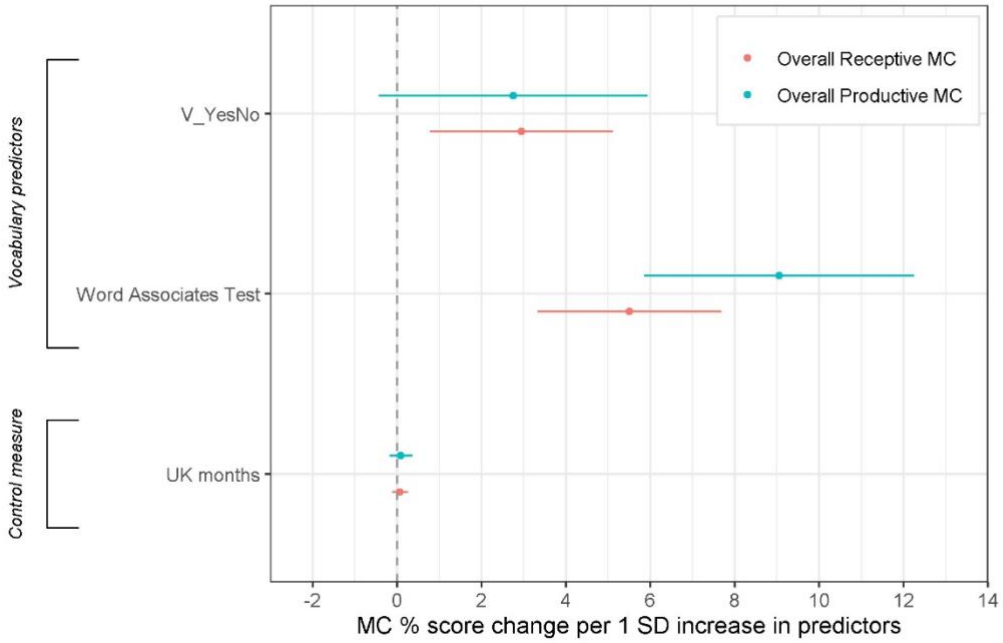


Figure 6: Models 1-2 standardised coefficients and 95% confidence intervals: criterion - Overall Receptive MC (Model 1), Overall Productive MC (Model 2); predictors – V_YesNo, WAT, UK months.

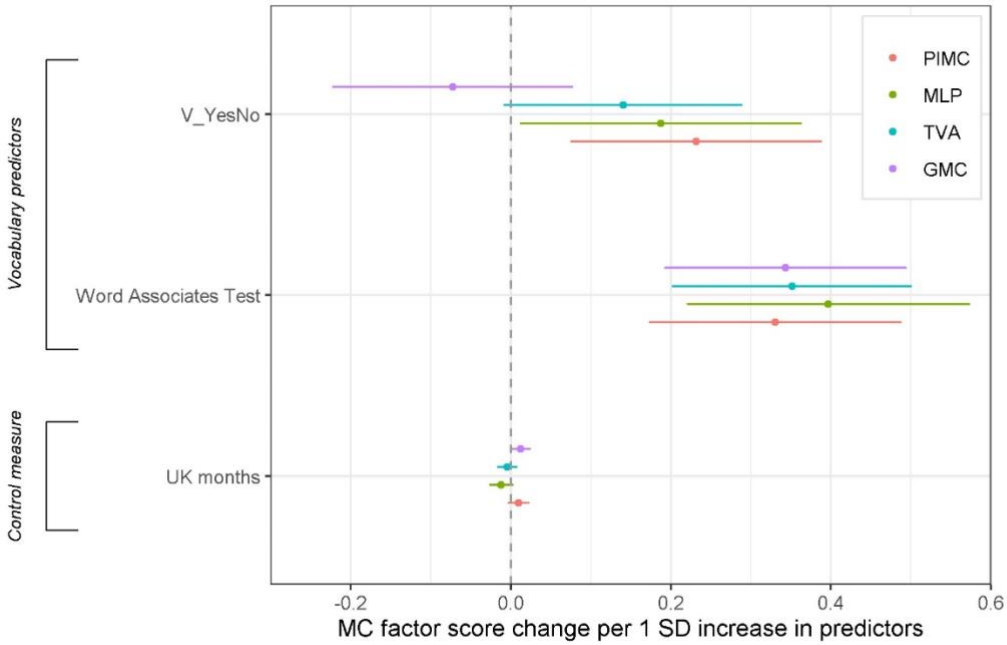


Figure 7: Models 3-6 standardised coefficients and 95% confidence intervals: criterion – Productive Illocutionary MC (Model 3), Metaphor Language Play (Model 4), Topic/Vehicle Acceptability (Model 5), Grammatical MC (Model 6); predictors – V_YesNo, WAT, UK months.

Post-hoc checks of variance inflation factors indicated no issue with collinearity, and normally distributed, homoscedastic residuals, with relatively equal variation across the range of the fitted values for all models.

5. Discussion

The present study sought to show how various types of L2 MC relate to two measures of general proficiency (RQ1) and vocabulary size and depth (RQ2) for this type of learner. While previous research and intuition suggest these constructs are generally all interrelated, the analytic techniques used allowed us to disentangle and isolate these relationships, thus quantifying unique and shared patterns of overlap. First (this section), we discuss the connections between the MC measures and two large scale standardised tests of general proficiency, characterising (as best we can) what these learners seem to be able to do with metaphor at the various OOPT and IELTS levels and how this information might be useful for language teachers and test/framework developers (Nacey 2013). Second (Section 5.2), we consider how closely the MC measures are intertwined with the types of vocabulary size and depth measured, and how associate thinking abilities (Littlemore and Low 2006a; Littlemore 2001; 2002; 2008) might help explain the connections discovered.

5.1 MC at different levels of general proficiency (RQ1)

The first research question used partial correlation analysis to identify the specific relationship between the MC measures and OOPT and IELTS after accounting for variance explained by the other proficiency measure (the one not in focus), the two vocabulary knowledge measures, and two sources of random, time-related variation. In each case, although the control variables played a role in explaining some of the original bivariate MC-proficiency relationships, the OOPT was found to share significant amounts of specific variance with Overall Receptive MC (14.3%), Overall Productive MC (10.7%), Productive illocutionary MC (8.6%), and Grammatical MC (10.8%); while IELTS shared significant amounts of specific variance with Overall Receptive MC (14.2%) and Topic/Vehicle Acceptability (8.7%). Other, weaker relationships potentially exist, but we focus here on those that are most clear and reliable.

So, what do these findings reveal about these learners' competence with metaphor at the various OOPT and IELTS levels? According to OOPT proficiency descriptors, C1 and C2 level learners are "likely to be able to use metaphor, idioms, and colloquialisms, to convey finer shades of meaning" (Purpura 2021). The present findings help provide a more specific understanding of this statement for this type of learner. At an OOPT indicated level of C1 or C2, roughly akin to IELTS 7/7.5, although IELTS-CEFR mapping is problematic (https://takeielts.britishcouncil.org/teach-ielts/test-information/scores-explained), these learners could mostly accurately recognise correct multiple-choice options and, to a lesser extent, supply gap-fill options for a wide range of metaphor types in relatively decontextualised, written-mode sentences/dialogues. Where the MC factors are concerned, the OOPT's alignment with Productive Illocutionary MC is probably explained by the pragmatic knowledge tapped by both measures, while its connection with Grammatical MC may be attributable to the OOPT's inclusion of test items involving phrasal verbs. In this view, like Grammatical MC, the OOPT taps a common competence with more conventional, fixed (rather than creative, flexible) language (see Table 1).

Importantly, learners at B2, B1, and even A2 levels seemed to recognise MC Test Battery metaphors with a moderate amount of accuracy, although they are more likely to struggle when producing metaphor within these types of item contexts. The clear implication for OOPT proficiency descriptor interpretation is that these learners (at least) do not seem to need to reach C1/C2 before they can recognise and even supply such metaphor with a degree of success, important points for language teachers to bear in mind when selecting, using and adapting MC Test Battery items as classroom/practice tasks, and where the OOPT is used to place students in streamed proficiency groups. Further, the findings are of relevance to researchers and practitioners interested in the development of metaphor in learner English, its intersection with the CEFR and other proficiency frameworks (Nacey 2013) and the continual development of the OOPT rubrics (Purpura 2004; 2021).

OOPT and IELTS have similar strength, individual connections with Overall Receptive MC, suggesting they both tap some kind of receptive MC that the other does not. However, the specific relationship between IELTS and Overall Productive MC is weak, implying that the longer, freer, written production and live speaking that IELTS elicits is not uniquely captured by Overall Productive MC, and likely develops more independently of it. This places a caveat on interpreting the scope of productive knowledge elicited by the MC Test Battery (O'Reilly and Marsden 2021). However, IELTS *does* have a unique connection with Topic/Vehicle Acceptability, whereas OOPT does not, perhaps because IELTS and Topic/Vehicle Acceptability both tap common knowledge and thinking related to semantic and syntactic associations, sensitivity to L1 norms and boundaries of acceptability. The role of such associations becomes even clearer as we turn to consider the relationships between MC and vocabulary knowledge.

5.2 MC and vocabulary knowledge (RQ2), a matter of associations?

The second research question investigated the specific relationships between vocabulary knowledge and MC, extending previous enquiries into vocabulary size and depth as combined and individual statistical predictors of language use (Qian 2002; Schmitt 2014) to the sparse MC literature on these connections (Azuma 2005). The multiple regression analyses showed that the three predictors (one entered for control purposes) accounted for just over one third of the variance in both Overall Receptive MC and Productive MC and between one fifth and one third in the MC factors. This implies the latter, uncovered by O'Reilly and Marsden (2021), were more conceptually independent from vocabulary knowledge than overall/composite receptive and productive MC. Unfortunately, there are no studies with which we can directly compare the MC-vocabulary knowledge relationships; Azuma (2005), for example, did not control for the vocabulary size-depth overlap when correlating these measures with MC, even so, her results were mixed. Interestingly, Qian (2002) found almost twice the degree of overlap than our models showed when individually correlating vocabulary size, depth (word association), and synonym recognition measures with TOEFL reading comprehension scores. His vocabulary knowledge measures explained even more criterion variance in combination than individually, suggesting a distinct explanatory role for size and depth (as Qian measured them) and that vocabulary knowledge seems to be more closely linked to a standardised reading comprehension test than MC (as we measured it).

In terms of size versus depth, the main finding for RQ2 was that the WAT explained everything about the MC measures that the V_YesNo did, and more, while the V_YesNo generally offered no unique explanatory power above and beyond the WAT. This seems to confirm that L2 MC draws on something other than, simply, a large vocabulary. (Reassuringly, months spent in the UK, our intended control measure in these analyses, had no unique predictive power and so is not discussed further).

So, why does the WAT have such a close connection to the MC Test Battery, and in particular, with Overall Productive MC and Metaphor Language Play? The answer, we think, has to do with the fact that these measures present stimuli that engage learners in a broad search and filtering of connected meanings and collocations to construct possible interpretations and productions, i.e., associative thinking (Carroll 1993; Littlemore and Low 2006a; Littlemore 2001; 2002; 2008). The WAT tests recognition of paradigmatic, syntagmatic, and analytic associates given a stimulus/headword (Section 3.2), an ability that goes hand in hand with generating metaphors throughout the MC Test Battery (Overall Productive MC) and items that involve recognising and producing creative and humorous extensions to re-literalised idioms and continuing a metaphorical discourse (Metaphor Language Play). In fact, with Metaphor Language Play items, test-takers were almost certainly encountering and producing new combinations of language and ideas, adapting existing knowledge to discourse situations with comparatively fewer pre-existing, formulaic, linguistic solutions (O'Reilly and Marsden 2021), i.e., engaging in "a type of active 'language play'" (Littlemore and Low 2006a: 56; see also Cook 1997; 2000; Crystal 1998).

The WAT also had a unique relationship with all other MC factors, which may suggest more than one type of associative thinking ability. For Productive Illocutionary MC, successful learners needed to generate associations that would help explain physical and natural world entities to children (Test 6-Heuristic-P) and similes/metaphors with appropriate emotional connotations (Test 7-Feelings-P). For Topic/Vehicle Acceptability, learners primarily needed to mentally search a source domain and evaluate whether the Vehicle term presented falls within acceptable semantic and syntactic boundaries (Test 3-Vehicle Acceptability-R). For example, *his blood began to boil as he started shouting* is an acceptable extension of the conceptual metaphor ANGER IS A HOT FLUID IN THE BODY, whereas *he bubbled as he began shouting* is not, deciphering this is helped if the learner knows (or guesses) that *blood* collocates with *boil* much more so than with *bubble* (Littlemore 2004c). For Grammatical MC, learners were given a specific meaning (e.g., *discouraged*) and required to recognise or retrieve suitable particles to complete metaphorical phrasal verbs (Test 1-Phrasal verbs-R and -P).

Seemingly, the WAT (itself a measure of receptive knowledge) shared a connection with both receptive and productive MC measures. Its particularly close connection with Overall Productive MC aligns with Henriksen's (1999) argument that greater depth of vocabulary knowledge is necessary for better productive control in general, extending this to controlled production of linguistic metaphor. However, we suspect the WAT's association with MC would not generalise equally well to all L1s, proficiencies, and/or learning contexts. Azuma (2005) found, in contrast to our results, that a simplified WAT version was anxiety inducing for her learners and had no significant correlation with a combined receptive and productive MC measure.

Vocabulary size was a comparatively poor predictor of L2 MC, somewhat surprising given its strong association with many language competences (for an early example involving the Eurocentres

Yes/No test, see Meara and Jones 1988; see also Schmitt 2010). This finding may be explained by the fact that V_YesNo elicits recognition of form, but with comparatively less requirement for test-takers to reflect on meaning (Henriksen 1999), in contrast to the six MC measures, which did require substantial engagement with meaning. Another issue is that learners' L1 may impact on the V_YesNo's effectiveness. Meara and Miralpeix (2016) cite studies showing that L1 French learners seem to perform differently to other learners, possibly due to the large number of French-English cognates, and that L1 Japanese speakers appear more reticent in declaring they 'know' a word. Whether this was a factor with our learners (L1 Mandarin) is a question for further research.

Finally, although we were unable to discuss causality under the current research design, it seems prudent to end the discussion by posing the question for future researchers: to what extent does (1) an increase in a learner's vocabulary depth lead to better productive control of metaphor, (2) engaging in L2 metaphor through figurative thinking activities, language play etc., provide sufficient conditions for learners to strengthen connections within their L2 lexical network, thus increasing their vocabulary depth? We suspect the answers depend on whether MC is operationalised via linguistic means (as in our study), or via an instrument (perhaps non-linguistic) that more directly taps the core cognitive construct, and the directionality of interest on what one chooses as their ultimate teaching/learning goal. Whatever the case, the extent and nature of the exact uni-/multi-directional causal pathways between MC constructs in observational studies such as ours is subject to confirmation via experimental designs with relevant control and manipulation (e.g., a metaphor/vocabulary depth-focused teaching intervention) and to test multiple pathways, even more robust analytic techniques than those used here (e.g., Structural Equation Modelling, see Hancock and Schoonen 2015).

6. Limitations

The present study reveals some interesting group trends for L1 Mandarin L2 English speakers at UK universities, although it is likely that the relationships vary considerably both between individuals and as a function of group characteristics such as the L1, culture, and age of participants. These avenues could be pursued in future research. It should also be remembered that our study tapped a specific kind of MC, namely elicited MC, using untimed reading and writing tasks, relatively decontextualised items, and with freedom for participants to engage various skills and/or knowledge types. The data would logically reflect both declarative and proceduralised knowledge (DeKeyser 2017a; 2017b; 2018), and both crystallised and fluid intelligence. However, we cannot be sure since we did not collect data on which items were novel for these test-takers, the sorts of knowledge and/or skills engaged during the test itself, or the learning mechanisms behind the knowledge/skills that learners had previously acquired and that they brought to the test. While these are questions for future research, we would not expect findings to necessarily generalise to naturalistic MC, where learners are interacting with real interlocutors in a dynamic, real-world setting.

Potential methodological limitations include the challenges of MC test development, described by O'Reilly and Marsden (2021), such as those related to construct operationalised and factor retention, and variation in test setting (despite the non-statistical difference). Although these may have influenced our findings to some extent, we do not think that the broad patterns observed would alter to a very great extent in the de-contextualised written mode. Other methodological limitations include our use of a single rather than multiple measures of vocabulary size and depth, due to the already heavy demands on our participants' time. While it is uncertain how other tests would perform, vocabulary size measures such as the Vocabulary Levels Test might also be poor predictors, given they measure a similar construct to the V_YesNo (Mochida and Harrington 2006).

For IELTS, participants reported the score from their most recent attempt rather than repeating this test for the purposes of the present study. While we were able to obtain estimates provided within the past two years for the majority of participants (considered valid by Higher Education institutions), some scores were older or unavailable, although we statistically controlled for this potential confound. Nonetheless, caution should be exercised when interpreting relationships involving this variable.

Finally, we also acknowledge that a longitudinal, mixed-setting approach might provide more ecologically valid evidence about process and MC development. For example, documenting day-to-day, real-world encounters with metaphor, alongside periodic measures of vocabulary knowledge, proficiency, and other relevant variables, which could provide rich evidence about the complexities of MC development.

7. Conclusion

The present study aimed to explore the relationships between L2 MC, vocabulary knowledge and general language proficiency, building on and adding nuance to previous enquiries. The MC-proficiency results showed that different kinds of MC share special relationships with two different proficiency measures, what these learners could do with metaphor at various proficiency levels, and how references to metaphor in proficiency descriptors might be more precisely interpreted. Importantly, the kinds of MC elicited in this study were not the privilege of CEFR C1/C2 learners but were demonstrated, to one degree or another, by participants at CEFR A2 and above. The MC-vocabulary knowledge relationships showed Read's (1993 1998) WAT was closely linked to elicited MC. The strong connection between productive MC and vocabulary depth extends Henriksen's (1999) proposed link between a more advanced lexical network and better productive control to the realm of productive L2 metaphor, while the MC factor-WAT relationship shows the centrality of different (but related) types of associative thinking in metaphor use and language learning more generally (Carroll 1993; Littlemore and Low 2006a; Littlemore 2001; 2002; 2008). However, vocabulary depth is about more than simply word association (Schmitt 2014), and so future research might explore the role that different depth

aspects (e.g., knowledge of paradigmatic/syntagmatic/analytic associates, polysemy, derivation) play in MC, individually and in combination.

The findings raise several points for pedagogical reflection. The clear association between vocabulary depth and MC (particularly productive recall and more playful aspects) suggests an important role for tasks that strengthen the lexical network with knowledge of synonyms and collocations, and room for guided experimentation with language and ideas. For specific ideas, we refer readers to Low's (1988) and Littlemore and Low's (2006a; 2006b) suggested activities for exploring the boundaries of metaphor, language play with metaphor, and other practising figurative thinking, i.e., doing more than merely task-essential vocabulary and phrase learning. For such activities, language teachers could utilise the items and response data from publicly available measures of MC (e.g., the MC Test Battery, O'Reilly and Marsden 2021). Existing instrumentation could also serve to help monitor learners' MC development and help teachers provide feedback on progress, although pedagogies should consider the relevance and transferability of the metaphors/aspects of MC to the particular variety of English being learned, and participants' language background, culture, cognitive maturity, and proficiency.

In high(er) stakes testing, the role of MC is less clear. While studies such as the current one emphasise the importance of MC in L2 learning/language use, we recognise that high stakes test developers face many challenges. It is therefore difficult to say how MC might become more of an embedded construct in language testing over the next few years, despite the many benefits offered.

Finally, on the methodological side, we have sought to advance the quantitative inquiry into L2 MC and its various correlates. The MC Test Battery and scoring procedures are fully transparent and our use of multiple linear regression analysis in line with developments in L2 MC literature (e.g., Chen and Lai 2015) and L2 research more generally (Plonsky and Oswald 2017). We hope that future studies will use our materials and develop more sophisticated models, in which pathways of causality, and their implications for L2 teaching, can be specified and tested.

8. References

Aleshtar, M. H., & H. Dowlatabadi. 2014. Metaphoric competence and language proficiency in the same boat. *Procedia - Social and Behavioral Sciences, 98*, 1895-1904. doi:10.1016/j.sbspro.2014.03.620

Allami, H., & J. Aghajari. 2014. Pragmatic knowledge of assessment in listening sections of IELTS tests. *Theory and Practice in Language Studies, 4*(2), 332-340.

Azuma, M. 2005. *Metaphorical competence in an EFL context*. Tokyo: Toshindo.

Bachman, L. F. 1990. *Fundamental considerations in language testing*. Oxford: University Press.

Benson, N. F., Beaujean, A. A., McGill, R. J., & S. C. Dombrowski. 2018. Revisiting Carroll's survey of factor-analytic studies: Implications for the clinical assessment of intelligence. *Psychological Assessment*, 30(8), 1028-1038.

Boers, F., Eyckmans, J., & H. Stengers. 2007. Presenting figurative idioms with a touch of etymology: More than mere mnemonics? *Language Teaching Research, 11*(1), 43-62. doi.org/10.1177/1362168806072460

Boers, F., Demecheleer, M. Coxhead, A., & S. Webb. 2014. Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research, 18*(1), 54-74. doi.org/10.1177/1362168813505389

Carroll, J. B. 1993. Human cognitive abilities: A survey of factor-analytic studies. Cambridge: Cambridge University Press

Chen, Y.-C., & L. H. Lai. 2015. Developing EFL learners' metaphoric competence through cognitive-oriented methods. *Iral-International Review Of Applied Linguistics In Language Teaching, 53*(4), 415-438. doi:10.1515/iral-2015-0019

Cohen, J. 1988. S*tatistical power analysis for the behavioral sciences*. New York: Academic Press.

Cook, G. 1997. Language play, language learning. *ELT Journal, 51*(3), 224.

Cook, G. 2000. *Language play, language learning*. Oxford: Oxford University Press.

Crystal, D. 1998. *Language play*. London, UK: Penguin.

DeKeyser, R. 2017a. Knowledge and skill in ISLA. In S. Loewen and M. Sato (eds.). *Routledge handbook of Instructed Second Language Acquisition* (pp. 15-32). London: Routledge

DeKeyser, R. 2017b. Age in learning and teaching grammar. In H. Nassaji (ed.). TESOL *Encyclopedia of English language teaching*. New York: Wiley.

DeKeyser, R. 2018. Task repetition for language learning: A perspective from skill acquisition theory. In M. Bygate (Ed.), *Learning language through task repetition* (pp. 27-41). Amsterdam (NL): Benjamins.

Farvardin, M. T., & M. Koosha. 2011. The role of vocabulary knowledge in Iranian EFL students' reading comprehension performance: Breadth or depth? *Theory and Practice in Language Studies, 1*, 1575–1580. doi:10.4304/tpls.1.11.1575-1580

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.

Hall, C. J. 2014. Moving beyond accuracy: From tests of English to tests of 'Englishing'. *ELT Journal, 68*(4), 376-385. doi:10.1093/elt/ccu016

Hancock, G.R., & R. Schoonen. 2015. Structural Equation Modeling: Possibilities for language learning researchers. *Language Learning, 65*(51), 160-184. doi:10.1111/lang.12116

Henriksen, B. 1999. Three dimensions of vocabulary development. *Studies in Second Language Acquisition, 21*, 303-317. doi:10.1017/S0272263199002089

Hoang, H. & F. Boers. 2018. Gauging the association of EFL learners' writing proficiency and their use of metaphorical language. *System, 74*, 1-8.

Johnson, J., & T. Rosano. 1993. Relation of cognitive style to metaphor interpretation and second language proficiency. *Applied psycholinguistics, 14*(2), 159-175.

Kövecses, Z. 2003. Language, figurative thought, and cross-cultural comparison. *Metaphor and Symbol, 18*(4), 311-320. doi:10.1207/S15327868MS1804_6

Lakoff, G., & M. Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

Landis, J. R., & G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

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

Linck, J.A., & I. Cunnings. 2015. The utility and application of mixed-effects models in second language research. *Language Learning, 65*(51), 185-207. doi:10.1111/lang.12117

Littlemore, J. 2001. Metaphoric competence: A language learning strength of students with a holistic cognitive style? *TESOL Quarterly, 35*(3), 459-491. doi:10.2307/3588031

Littlemore, J. 2002. Developing metaphor interpretation strategies for students of economics: a case study. *Les Cahiers de l'APLIUT, 22*(4), 40-60.

Littlemore, J. 2004a. Interpreting metaphors in the language classroom. *Les Cahiers de l'APLIUT 23*(2) 57-70.

Littlemore, J. 2004b. The effect of cognitive style on vocabulary learning strategy preferences. *Iberica, The Academic Journal of AELFE 7*, 5-32.

Littlemore, J. 2004c. What kind of training is required to help language students use metaphor-based strategies to work out the meaning of new vocabulary? *Documentao de Estudos em Linguistica Teorica e Aplicada DELTA 20*(2): 265-279.

Littlemore, J. 2008. The relationship between associative thinking, analogical reasoning, image formation and metaphoric extension strategies. In M. Zanotto, L. Cameron and M. Cavalcanti (eds.), *Confronting Metaphor in Use: An Applied Linguistic Approach* (pp. 199-222). Amsterdam/Philadelphia: John Benjamins.

Littlemore, J. 2010. Metaphoric competence in the first and second language: similarities and differences. M. Putz and L. Sicola (eds.) *Cognitive processing in second language acquisition: Series converging evidence in language and communication research* (pp. 293-316). Amsterdam: John Benjamins.

Littlemore, J., Chen, P., Koester, A. and J. Barnden. 2011. Difficulties in metaphor comprehension faced by international students whose first language is not English. *Applied Linguistics, 32*(4), 408-429.

Littlemore, J., Krennmayr, T., Turner, J., and S. Turner. 2014. An Investigation into Metaphor Use at Different Levels of Second Language Writing. *Applied Linguistics, 35*(2), 117–144.

Littlemore, J., & G. D. Low. 2006a. *Figurative thinking and foreign language learning*. Basingstoke: Palgrave Macmillan.

Littlemore, J., & G. D. Low. 2006b. Metaphoric competence, second language learning, and communicative language ability. *Applied Linguistics, 27*(2), 268-294. doi:10.1093/applin/aml004

Low, G. D. 1988. On teaching metaphor. *Applied Linguistics, 9*(2), 125-147. doi:10.1093/applin/9.2.125

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*, 23, pp. 155-173).

Marsden, E., Mackey, A., & L. Plonsky. 2016. The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages*: Routledge, pp. 1-21. doi.org/10.4324/9780203489666

McManus, K., & E. Marsden. 2018. Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics, 40*(1), 205–234.

McNeish, D. 2018. Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412-433. doi:10.1037/met0000144

Meara, P., & B. Buxton. 1987. An alternative to multiple choice vocabulary tests. *Language Testing, 4*(2), 142-154. doi:10.1177/026553228700400202

Meara, P., & G. Jones. 1988. Vocabulary size as a placement indicator. In P. Grunwell (ed.), *Applied Linguistics in Society* (pp. 80-87). London: CILT.

Meara, P., & I. Miralpeix. 2015. V_YesNo v1.0. Retrieved from www.lognostics.co.uk/

Meara, P., & I. Miralpeix. 2016. *Tools for researching vocabulary*. Bristol: Multilingual Matters.

Meara, P., & B. Wolter. 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.

Mehrpour, S., Razmjoo, S. A., & P. Kian. 2010. The relationship between depth and breadth of vocabulary knowledge and reading comprehension among Iranian EFL learners. *Journal of English Language Teaching and Learning, 53*(222).

Miles, J. N. V., & M. Shevlin. 2001. *Applying regression and correlation: a guide for students and researchers*. London: Sage.

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

Munzel, U., & E. Brunner. 2000. Nonparametric tests in the unbalanced multivariate one-way design. *Biometrical Journal, 42*(7), 837-854. doi:10.1002/1521-4036(200011)42:7%3C837::AID-BIMJ837%3E3.0.CO;2-S

Nacey, S. 2013. *Metaphor in learning English*. Amsterdam: John Benjamins.

Nation, I. S. P. 1983. Testing and teaching vocabulary. *Guidelines, 5*(1), 12-25. doi:10.1002/9781444315783.ch28

NourMohamadi, E. 2010. *Conceptual metaphor and the acquisition of English metaphorical competence by Persian English majors: A cognitive linguistic approach*. Unpublished PhD thesis. Allame Tabataba'i University, Iran.

O'Reilly, D. 2017. *An investigation into metaphoric competence in the L2: A linguistic approach*. Unpublished PhD thesis. University of York, UK.

O'Reilly, D., & E. Marsden. 2021. Eliciting and measuring L2 metaphoric competence: Three decades on from Low (1988). *Applied Linguistics, 42*(1), 24-59, doi.org/10.1093/applin/amz066

Pellicer-Sánchez, A., & N. Schmitt. 2012. Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing, 29*(4), 489-509. doi:10.1177/0265532212438053

Perneger, T. 1998. What's wrong with Bonferroni adjustments? *BMJ, 316*. doi:10.1136/bmj.316.7139.1236

Plonsky, L., & F. L. Oswald. 2017. Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition, 39*, 579-592. doi:10.1017/S0272263116000231

Pollitt, A. 2021. The Oxford Online Placement Test: The meaning of OOPT scores. Retrieved from http://fdslive.oup.com/www.oup.com/elt/feature/assessment/oxford-placement-test-what-does-it-measure.pdf?cc=gb&selLanguage=en&mode=hub

Purpura, J. 2004. *Assessing grammar*. Cambridge: Cambridge University Press.

Purpura, J. 2021. The Oxford Online Placement Test: What does it measure and how? Retrieved from http://fdslive.oup.com/www.oup.com/elt/feature/assessment/oxford-placement-test-the-meaning-of-opt-scores.pdf?cc=gb&selLanguage=en&mode=hub

Qi, G.-Y. 2016. The importance of English in primary school education in China: perceptions of students. *Multilingual Education, 6*(1), 1-18. doi:10.1186/s13616-016-0026-0

Qian, D. D. 1999. Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*, 282–307. doi:10.3138/cmlr.56.2.282

Qian, D. D. 2002. Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536. doi:10.1111/1467-9922.00193

R Core Team. 2019. *R: A language and environment for statistical computing (R Foundation for Statistical Computing)*. Retrieved from https://www.R-project.org/

Read, J. 1993. The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355-371. doi:10.1177/026553229301000308

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.

Richard, J.-P. J. 2011. Does size matter? The relationship between vocabulary breadth and depth. *Sophia International Review, 33*, 107-120.

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. doi:10.1111/lang.12077

Schmitt, N., Ng, J., & J. Garras. 2011. The Word Associates Format: Validation evidence. *Language Testing, 28*(1), 105-126. doi:10.1177/0265532210373605

Schmitt, N., Schmitt, D., & C. Clapham. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88. doi:10.1177/026553220101800103

Steen, G., Dorst, A. G., Berinke Herrmann, J., Kaal, A. A., Krennmayr, T., & T. Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam: John Benjamins.

Thurstone, L. L. 1935. *The vectors of mind*: University of Chicago Press.

Trenkic, D., & M. Warmington. 2019. Language and literacy skills of home and international university students: How different are they, and does it matter? *Bilingualism: Language and Cognition, 22*, 249-365. doi:10.1017/s136672891700075x

Vermeer, A. 2001. Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217-234. doi:10.1017/S0142716401002041

Winter, B. 2020. *Statistics for linguistics: An introduction using R*. Abingdon: Routledge.

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.

Wesche, M. B., & T. S. Paribakht. 1996. Assessing second language vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review, 53*(1), 13-40. doi:10.3138/cmlr.53.1.13

Zhao, Q., Yu, L., & Y. Yang. 2014. Correlation between receptive metaphoric competence and reading proficiency. *English Language Teaching, 7*(11), 168-181. doi:10.5539/elt.v7n11p168