Applied Linguistics 2021: 42/1: 24–59
doi:10.1093/applin/amz066 Advance Access published on 25 January 2020

Eliciting and Measuring L2 Metaphoric Competence: Three Decades on from Low (1988)

*DAVID O'REILLY and EMMA MARSDEN

Department of Education, University of York
*E-mail: david.oreilly@york.ac.uk

In 1988 and 2006, *Applied Linguistics* published Low's and Littlemore and Low's seminal theoretical accounts of second language (L2) metaphoric competence (MC). Meanwhile, attempts to elicit metaphor-related skills/competences have been mixed. Instrumentation has varied in reliability, been limited in scope, and used arguably flawed reliability coefficients (McNeish 2018). Factor analysis, used in first language (L1) MC and other areas of L2 research (Loewen *et al.* 2009; Plonsky and Gonulal 2015), has not been used to explore latent L2 MC variables. To address these issues, we developed a large battery of MC tests to elicit Low/Littlemore's constructs, administering it to 112 L1 Mandarin speakers of L2 English and 31 L1 English speakers. Data cleaning revealed some operationalization challenges, but resulting overall reliability was high and demonstrated innovative use of ordinal omega as a powerful alternative to Cronbach's alpha, the most common instrument reliability index in L2 research (Plonsky and Derrick 2016). Exploratory factor analysis suggested four latent L2 MC variables: Productive Illocutionary MC, Metaphor Language Play, Topic/Vehicle Acceptability, and Grammatical MC, broadly supporting Low/Littlemore's proposals.

INTRODUCTION

Research into how people comprehend and produce metaphor, broadly speaking their metaphoric competence (MC), has existed since early studies on first language (L1) child, adolescent and adult acquisition, and use of figurative language (Pollio and Burns 1977; Pollio and Smith 1979, 1980). Low's (1988) proposal of several metaphor-related skills as characteristics of MC was the first major discussion on the importance of metaphor for second/foreign language (L2) learners and teachers. For example, Low described the need for learners to develop aspects of MC such as: a knowledge of the boundaries of conventional metaphor (e.g. 'The theory needs a better framework/greater support/a firmer foundation' as conventionally acceptable (p. 130)); the ability to interpret and control 'hedges' (e.g. 'literally' in the metaphor 'He literally hit the roof, he was so angry' (p. 133)); and the awareness of socially sensitive metaphors (pp. 133–4) (e.g. animal metaphors used to describe undesirable human habits or attributes).

In their 2006 book and article, Littlemore and Low developed Low’s (1988) work to argue for the relevance of metaphor and figurative thinking to grammatical, textual, illocutionary, and sociolinguistic components of language competence and strategic competence (Bachman 1990). For example, learners not only need to be able to comprehend and produce figurative idioms, but to be able to playfully re-literalize them in acceptable ways, as in ‘I’ve been sitting on the fence so long my bottom is beginning to hurt’ (2006a: 130). This process is thought to cause learning via destabilization of the interlanguage system (MacArthur 2010). While informal experiences were reported of learners engaging in the skills proposed by Littlemore and Low (2006a: 129–32), the work had theoretical rather than empirical ambitions.

Since the publication of these works in 1988 and 2006, the development and application of reliable metaphor identification procedures (MIPs) have revealed that a substantial proportion of language is metaphorical, in English, for example, 17.5 per cent of words used in academic discourse, 15.3 per cent in news texts, 10.8 per cent in fiction, and 6.8 per cent in spoken conversation (Steen *et al.* 2010: 194–208). These procedures have also facilitated research into metaphor use in important, real-world contexts, including (recently) UK secondary school talk about climate change (Deignan *et al.* 2019) and palliative healthcare (Semino *et al.* 2018). In short, there is now even more support for Low’s (1988) general proposition that metaphor plays a central role in human language, and as such, is relevant to virtually all second/foreign language learners and should be given prominence in language education.

However, the metaphor-related skills/sub-competences theorized in Low/Littlemore’s studies have never driven the development of tests of L2 MC, meaning their measurability as constructs is unknown. Existing L2 MC tests exhibit problems of mixed reliability and the need for substantial piloting and item refinement. Furthermore, factor analytic methods, used to investigate the theorized conceptual structure of L1 MC, have not been applied in L2 MC research. We devote the current study to addressing these issues through the design and analysis of a large battery of theory-driven L2 MC tests. Such robust instrument development is essential so that reliable measures can then be used to explore the MC construct further to investigate, for example, the extent to which MC relates to vocabulary knowledge and general language proficiency (O’Reilly and Marsden 2020a). Whilst considerable theoretical and empirical work has focused on reliably eliciting other aspects of language competence such as morphosyntax, the lexicon, pragmatics, and phonology, very little work has systematically focused on eliciting a theory-driven measurement of MC. By grounding our study in research from a variety of fields—linguistics (linguistic metaphor identification), second language acquisition (SLA, language competence theory), psychometrics (instrument development, construct measurement)—we hope to advance and contribute to an innovative, transdisciplinary, and multilingual MC research agenda (The Douglas Fir Group 2016).

We begin the current study by considering what MC is, challenges concerning its reliable measurement, and approaches to modelling its underlying structure. Next, we set out our aims to elicit and reliably measure subcomponents of L2 MC and identify factors underlying the MC Test Battery scores. These are followed by a detailed methodology reporting the development of the MC Test Battery, main results, and discussion of them. After presenting several limitations and future directions we conclude by emphasizing the study's main contributions and applications.

LITERATURE REVIEW

Operationalizing MC: definitions and key concepts

There is continued debate over what, exactly, counts as metaphor. For our purposes, we adopt Low's working definition of metaphor as '. . . a reclassification which involves: Treating X as if it were, in some ways, Y' (1988: 126) as a straightforward way of referring to the wealth of figurative language (also including metonymy, simile, idiom, etc.) that learners need to deal with.

Researchers often distinguish between linguistic (language-based) metaphor, for example, 'the driver was . . . fuming/going to explode/hot under the collar', and conceptual (thought-based) metaphor, here ANGER IS FIRE (Lakoff and Johnson 1980). In language, metaphor is described in terms of the topic conveyed (anger) and vehicle/words used (e.g. 'fuming'), whereas conceptually, the distinction is described as a target domain (ANGER) and a source domain (FIRE). (For an overview of Conceptual Metaphor Theory, see Gibbs 2014.) In the current study, we employ both linguistic and conceptual routes to operationalizing metaphor in the development of the MC Test Battery, since both are relevant to L2 learners' MC.

MC (L1, L2, L3, etc.) involves the comprehension, production, awareness, and retention of metaphor in speaking, writing, reading, and/or listening. In order to identify metaphors in written or spoken production, methods have been developed, such as the MIP (Pragglejaz 2007) and its later refinement the MIP VU University Amsterdam (MIPVU) (Steen *et al.* 2010), that classify a lexical unit (e.g. a word or cluster of words) as metaphorical if it is understood by contrasting a discourse- (or context-) dependent meaning with a (more) 'basic' (or concrete) meaning.

Importantly, however, MC does not equate only to *production* of metaphor-related words and phrases that can be measured by identification procedures such as the MIP and MIPVU. Rather than concerning simply the quantity of metaphor produced (whereby increasing competence is reflected in the production of 'more' metaphors), MC is in fact a multi-dimensional set of skills and sub-competences. These sub-competences include the ability to use metaphorical language appropriately, strategically, and playfully, in order to engage in real-world interactions. For example, a metaphor can continue throughout an entire conversation (Low 1988: 134–35), or a metaphorical idiom/proverb/saying (e.g. ‘there’s no use crying over spilled milk!’) can be invoked to summarize the moral of an experience or signify a desire to change topic (Littlemore and Low 2006a: 144–46).

To investigate MC, researchers use ‘naturalistic’ data, for example, unprompted spoken production (Pitzl 2016) and written assignments (Kathpalia and Carmel 2011; Nacey 2013), as well as ‘elicitation’ methods (e.g. comprehension/production tests, experimental stimuli). Compared with elicitation, naturalistic data are thought to offer better insight into authentic, spontaneous, language in use, but they present a less clear picture of the ‘boundaries’ of knowledge, that is, what is *not* known. For example, if a learner does not produce a particular metaphor or type of metaphor, the researcher cannot know whether the learner does not know it or just did not have occasion or desire to use it. Of course, very often it is difficult to draw a clear separation between naturalistic and elicited data as the two can be seen to overlap, such as in a role-play job interview eliciting spontaneous productions or a written essay eliciting pre-prepared content.

Elicitation of metaphor has been undertaken by a number of studies to date. One set (e.g. Littlemore 2001; Azuma 2005, various by Frank Boers and colleagues) has used comprehension and production tests with experimental stimuli that target very specific components of linguistic, conceptual or visual metaphors, or functions of metaphor. A small number of other studies have sought to develop tests measuring fluency, accuracy, and originality (creativity) of L2 metaphor use. Littlemore (2001), for instance, operationalized MC via computer-based rating tasks (measuring, e.g., the speed in finding meaning in metaphor) and pen and paper tests (measuring the originality of metaphor production). These tests tapped into the dynamic process of metaphor understanding and individual differences, such as cognitive styles, associated with metaphor production. However, Littlemore noted that she used a definition of MC ‘narrower than that proposed by Low (1988), who includes aspects of crystallized intelligence’ (2001: 461). Thus, her elicitation methods focused more on fluid mental processes involved in metaphor comprehension and production than the kinds of skills and concrete language and cultural knowledge that Low described, which form part of the basis for the tests developed in the current study.

In other metaphor elicitation studies, the pedagogical context seems to have driven test development. Azuma (2005) created tests of literal/figurative sentence writing using the frame *an/the X is a(n) adjective Y*, and literal/figurative understanding and use of idioms and proverbs sourced from specialized dictionaries. Azuma is one of few researchers to have developed tests argued to be context-sensitive, as they promoted polysemy skills that, specifically, Japanese learners of English as a foreign language (EFL) find challenging. The tests were also designed to reduce test-taker anxiety, thought to occur in this particular context (see also Zhao *et al.* 2014, who administered two of Azuma’s MC tests to Chinese EFL learners).

We found only one empirical study, by Kathpalia and Carmel (2011), in which Low/Littlemore's constructs were explicitly targeted. Using a more naturalistic approach than the aforementioned studies, these authors identified linguistic metaphors in relation to grammatical, illocutionary, textual, and sociolinguistic competences (Littlemore and Low 2006a,b) in 113 samples of essay text produced by Singaporean learners of L2 English. They found grammatical miscollocations to be the most frequent type of problem, present in 88 per cent of the scripts. These were mostly noun phrase miscollocations such as 'more researches [more *research*]', or verb–noun miscollocations such as 'keep our hobbies [*indulge in* our hobbies]' (p. 280). Illocutionary competence problems, concerning clichéd expressions, misused metaphor, and misplaced humour, occurred in 84 per cent of scripts, while textual competence problems (e.g. inappropriate metaphor signalling) occurred in 62 per cent, and sociolinguistic competence problems (e.g. inappropriate cultural references) in 19 per cent. The authors suggested that while the learners made numerous attempts to use metaphor in written production, they often appear to have lacked appropriate pre-fabricated language.

However, as Kathpalia and Carmel acknowledged, considerable caution should be exercised when operationalizing and applying a notion of 'appropriate language' or 'correctness' to learner productions. In Singapore, English is an official, and therefore Second (rather than Foreign) Language for L2 learners. Given Singapore's linguistic diversity, English is also a lingua franca for communication. While the authors specified 'standard English' (p. 278) as a criterion for determining 'target' (i.e. deemed 'appropriate') versus 'interlanguage' (deemed 'inappropriate') forms, they also clearly explain that many collocations of the latter type would be acceptable in local varieties of English, and that the 'target' forms specified were not intended as prescriptive or superior to other world English or lingua franca forms.

Concerning the latter, lingua franca contexts offer rich examples of metaphor use, where MC is often characterized by speakers drawing from multilingual resource pools and knowledge of interlocutors. Pitzl (2016), for example, discussed an L1 Dutch speaker's production of 'put my hands into the fire for it' (p. 301) when interacting with L1 German speakers, a creative coinage in English, which would have made particular sense to the interlocutors on account of similar metaphors in Dutch and German.

The above examples highlight the need for researchers to explain the approach to 'correctness' taken and to reflect on its usefulness not only to help operationalize MC, but also for the specific learners and contexts studied. They also raise the fundamental question as to whether 'target' linguistic metaphor forms and functions can exist, and if so, whether they should be learnt/taught.[1] (We note that this question applies to all aspects of language learning, but for necessity, the discussion here is limited to MC.)

To position ourselves on the first issue, we believe that trying to develop one's MC means, inevitably, *aiming* for something. It is, therefore, difficult to see how this process would not involve the targeting of specific linguistic metaphors, either directly or indirectly (e.g. as a by-product of trying to master the ability to perform a certain pragmatic function using metaphor). While target metaphors in this sense exist, the large number of varieties of English, discourse domains, registers, possible combinations of interlocutors, etc., make it difficult to know *which* particular metaphors, if any, might be used to foster the development of MC in any given situation and, for our purposes, selected to measure learners’ competence. Any generalized approach to scoring comprehension and production for correctness is further complicated by that fact that most, if not all, metaphors are likely to differ greatly in their degree of specificity to a particular language variety, domain, genre, etc. Thus, a linguistic metaphor deemed to be, or contain, an error in one context due to unconventional wording, an unusual idea, etc., might be seen as a creative innovation in another.

On the question of whether target metaphors *should* be taught and indeed tested, our position closely aligns with Hall’s (2014) arguments about assessment more generally. We agree that MC tests should be developed to focus on what learners can do with English (in Hall’s terms ‘Englishing’) rather than use of any particular type of English, noting also that Low (1988) defined ‘competence’ in terms of ability to function socially rather than knowledge of a particular set of linguistic metaphors. Where appropriate, we take this approach to the tests developed in the current study. For example, for the majority of productive tests, the scoring criteria emphasize task completion and demonstration of a particular ability, with little to no stipulation about how this is done linguistically and no penalty for spelling/grammar ‘errors’ provided meaning is decipherable.

At the same time, and again in line with Hall (2014), we caution that adopting a more plurilithic view towards assessment does not mean discarding both baby and bathwater. Clearly, the use of stricter linguistic criteria (e.g. eliciting particular linguistic metaphors), and fostering the ability to confirm to a particular English variety has both a utilitarian purpose (e.g. helping learners operate within specific genres such as academic writing, preparation for particular employment) and a personal importance to many learners, whose aspirations to their chosen set of perceived norms should be respected. In our tests, stricter linguistic criteria became more important where it was necessary to elicit test-takers’ comprehension of specific linguistic metaphors (e.g. when testing receptive knowledge) in order to operationalize a particular aspect of MC as described by Low/Littlemore. We return to these issues, and their challenges, later in the article.

In sum, and despite the complexities of coding metaphor for ‘correctness’ or ‘appropriacy’, the studies cited thus far indicate that some of the competences theorized by Low/Littlemore *can* occur in L2 writing (Kathpalia and Carmel 2011) and multilingual spoken interaction (Pitzl 2016). However, the chief limitation of naturalistic data is the lack of information provided on the extent and boundaries of learners’ knowledge and skills on a wider range of MC subcompetences. For this, it is necessary to establish further means of *eliciting* metaphor in comprehension and production and measuring it reliably.

MEASURING MC

Reliability in MC research

Within some domains, measurement reliability has been reviewed in order to systematically examine methodological and reporting practices and to estimate overall reliability and related effect sizes, and their variability across study, instrument type, and participant features. Such systematic reviews have appeared in SLA research, both the wider field (Plonsky and Derrick 2016) and sub-domains such as L2 judgement tests (Plonsky *et al.* 2019) and L2 pronunciation (Saito and Plonsky 2019) and in the psychology literature (Rodriquez and Maeda 2006; Wheeler *et al.* 2011). However, within the subdomain of MC research, the extent of measurement reliability remains unclear. To investigate this, prior to our main study, we systematically documented the measurement and reporting of instrument, inter-rater, and/or intra-rater reliability in 33 empirical MC studies. These studies were found by searching for one of 'metaphor*' OR 'metonymy*' OR 'simile*' OR 'idiom*' OR 'figurative' OR 'analogy' AND one of 'competence' OR 'proficiency' OR 'learning' OR 'second language' OR 'foreign language' on the databases *Language and Linguistic Behaviour Abstracts* and *Educational Resources Information Center*, and backward searching through reference sections in journal articles and books.

The 33 studies (see Supplementary Material for full list), which dated 1970–2015, comprised empirical research published in peer-reviewed journals, books, conference proceedings, and PhD theses. A total of 18 studies had L1 participants only, 12 had L2 participants only, and three had both L1 and L2 participants. Importantly, with the exception of Kathpalia and Carmel (2011, see pages 5–6 above), no study attempted to operationalize Low's (1988) and Littlemore and Low's (2006a, b) proposed metaphor-related skills/sub-competences, and so we note that our review of measurement reliability reported below does not relate to measuring these authors' L2 MC models.

As we required a coding scheme suitable for secondary data about reliability, we used the scheme developed by Plonsky and Derrick (2016) for meta-analysing reliability in SLA, and adapted it for measuring MC. This scheme allowed us to record study, instrument, and participant features in our sample of studies. The first author coded all 33 studies, while a second rater (also an applied linguist) coded a subsample (>20 per cent of the data, $k = 7$ studies, consisting of 759 decisions involving nominal, ordinal, and continuous data). The agreement rate was 95 per cent. Disagreements were easily resolved through discussion and identified no systematic coding problems.

In these 33 studies, 176 administrations of MC instruments[2] (1 to 16 per study) were found. We examined both instrument and inter-/intra-rater reliability. Instrument reliability was reported in just 43 of these administrations (some of which were repetitions of the same test), across 10 of the studies, with Cronbach's alpha (α) the most common index, used in over half ($k = 25$), and Coefficient *H*, Kuder–Richardson, split-half, and Spearman–Brown methods also used, each in at least two applications. The number of test items ranged from 1 to 90, and sample sizes from 6 to 149. The median (Mdn) MC instrument reliability (0.76) suggested that MC tests may, on average, be less reliable than instruments used in SLA research generally (where the Mdn was found to be 0.82 from 1323 coefficients, Plonsky and Derrick 2016). Furthermore, while most MC instrument estimates were relatively close to the median (interquartile range/IQR = 0.67–0.82), 10 and 11 estimates lay below and above this range, respectively, with rather extreme minimum and maximum values (0.31–0.90, both α), suggesting substantial variation in instrument reliability.

Inter-rater reliability for scoring metaphor in written and spoken production was reported in 49 of a possible 111 administrations of scoring procedures, with a median of 0.82, again lower than in SLA (Mdn = 0.92, from 861 coefficients, Plonsky and Derrick 2016). Percentage agreement was the most commonly reported indication of inter-rater reliability, an indicator that does not compensate for chance agreements, and so many of these indices may be inflated. In only 26 of 47 cases where disagreements occurred, were these followed by a final, revised score after discussion. Intra-rater reliability estimates (the same person re-scoring the same data), potentially reportable in 111 administrations, were never provided.

In sum, not only has the reporting of reliability been patchy (in line with similar observations across other methodological syntheses, e.g. Plonsky *et al.* 2019), but, where it is reported, it has been somewhat lower than hoped. One likely reason for the underreporting of reliability in L1 MC research, specifically, is that most of the L1 studies in our list were conducted in the 1970s or 1980s, when technology for statistical computation was less available. An additional reason, for both L1 and L2 MC research, is that expectations about test development practices and reporting procedures inevitably evolve over time.

Indeed, even today there is scope for improvement in the choice of instrument reliability coefficients. While Cronbach’s alpha has played a central role in MC test reliability estimation and refinement, and indeed in all L2 research, its assumptions are rarely, if ever, met or (where possible) corrected for. McNeish (2018) suggests that Cronbach’s alpha requires that items contribute equally to the total scale score (tau equivalence), are continuous with normal distributions, measure the same construct (unidimensionality), and are not correlated via any other sources (uncorrelated errors). McNeish then proposed three conceptually similar alternatives—omega coefficients, coefficient *H*, and greatest lower bound—which determine instrument reliability via factor analysis of item loadings on a single latent dimension, to different degrees, bypassing alpha’s assumptions. Using empirical examples, McNeish demonstrates how (i) when the assumptions of Cronbach’s alpha are not met, using this coefficient can make scales appear less reliable and (ii) even when its assumptions are met, the alternatives yield justifiably higher values. (See, however, Raykov and Marcoulides (2019) who argued for the continued use of

Cronbach's alpha under appropriate conditions, challenging McNeish's reading of Cronbach's (1951) original assumptions.)

Our synthesis of the 33 studies reported above showed that none of these more appropriate coefficients had been used in L2 MC research, and only one, Coefficient *H*, in an L1 MC study. Thus, our approach to reliability (see section 'Analysis') addresses McNeish's and others' calls for the use of superior alternatives to Cronbach's alpha, a hitherto much-loved coefficient.

Other challenges for L2 MC measurement research

Possibly the biggest challenge facing researchers in this area is the relative lack of availability of piloted, validated, high-reliability materials used in peer-reviewed research, that tap a broad range of metaphor-related skills/sub-competences and knowledge. To illustrate, there are currently just 14 instruments or scoring/coding materials for measuring MC available on www.iris-database.org (Marsden *et al.* 2016), compared with 217 grammaticality/acceptability judgment tests (non-MC), 124 attitudinal questionnaires, and 44 working memory tests.

Other methodological shortcomings of MC tests include the use of dated, literary idioms as stimuli, such as 'the rotten apple injures its neighbours' or non-conventional (corrupt) forms, such as 'you cannot eat your cake and have your cake' (Azuma 2005, cf. Macmillan English Dictionary 2019).

MC research using (more naturalistic) production data is additionally challenged by the need to clarify and justify coding and scoring decisions. For example, in Kathpalia and Carmel (2011), while the authors defined metaphor as 'experiencing one concept in terms of another', the protocol for counting and classifying linguistic metaphors into the various categories is somewhat unclear, since a robust procedure (e.g. the MIP, Pragglejaz 2007) was not used, and reliability checks, if conducted, were not reported. The omission of these steps may, in part, explain a small number of questionable coding decisions, such as 'my fellow friends' being classified as an 'interim' (i.e. interlanguage) phrase, but 'my fellow mates' as a 'target' phrase.

More generally, identifying the effects of participant characteristics on L2 MC is also a challenge. For example, experimental studies have sometimes used fairly small groups of L2 learners with a wide range of L1s, which, given the cross-linguistic and cross-cultural differences in metaphor use (for an overview, see Kövecses 2010), is likely to account for some observations about L2 MC, but these cannot be examined systematically if there are very low numbers of learners with each L1. Proficiency categorization is also rarely adequately measured or reported (e.g. how 'high' and 'low' proficiency groups were determined in Aleshtar and Dowlatabadi 2014). Of course, these challenges are not unique to MC research, but apply to L2 research more widely (Thomas 2006).

A more conceptual challenge concerns the rationale behind some L2 MC research and the consequences this has had on the operationalization of the construct. For example, studies focusing on the teaching of L2 MC have tended to focus on a small number of forms or functions, such as the benefits of etymological elaboration for learning sets of L2 idioms. (See Boers *et al.* (2007) for calls to broaden the scope of this research.) A consequence of this is that the measures developed for this set of instruction studies have essentially been achievement tests eliciting limited subsets or unique features of MC.

Exploring underlying competences in L2 MC research

L2 researchers have yet to tap into methods frequently used in L1 MC research, such as factor analysis, for exploring latent MC dimensions and their psychometric properties. For example, in an L1 context, Beaty and Silvia (2013) found that timed odd-one-out letter sets, series completion, and visual-spatial reasoning tasks (fluid intelligence) and three timed verbal fluency tasks (broad-retrieval ability) best predicted the quality of creative metaphors, operationalized using a task in which participants described two past experiences using metaphor. Responses were scored on a 1–5 creativity scale by three raters trained to consider ‘remoteness’ (the conceptual distance of the Topic and Vehicle), ‘novelty’ (the originality of the response, such that clichés and common idioms received a low score), and ‘cleverness’ (the degree to which the response was funny, witty, or interesting). Conventional metaphor generation, on the other hand, was best predicted by vocabulary and general knowledge tasks (crystallized intelligence) and an openness to new experience measure. However, the extent to which different dimensions are quantifiable and reliably observable in L2 MC remains unexplored, a gap that the current study begins to address.

The present study

Thus far, L2 research has operationalized the MC construct in a variety of ways, but these exhibit numerous limitations and have largely neglected the potential of Low/Littlemore’s metaphor-related skills/sub-competences for both theory-building and, by extension, for informing assessment and pedagogical interventions. Investigation into the MC construct, such as its relationship to other aspects of L2 knowledge (e.g. vocabulary and general proficiency, see O’Reilly and Marsden 2020a) requires instrumentation that is theory-driven, valid, and reliable. However, reliability estimates in (L1 and L2) MC research have generally been underreported, lacking in rigour, and lower than the SLA field more generally. Also, elicitation methods have not drawn on the established theory about L2 MC contained within Low’s (1988) and Littlemore and Low’s (2006a,b) well-articulated, comprehensive, and highly cited metaphor-related skills/sub-competences. Moreover, the limitations of Cronbach’s alpha under certain conditions and the availability of superior alternatives indicate the need for a new approach to estimating instrument reliability, which we adopt here, to the best of our knowledge, for the first time in L2 research. Finally, there is no empirical research to date on eliciting and measuring the *underlying* structures of L2 MC.

Focusing on these gaps, the current study addressed two research questions:

> RQ1: To what extent can L2 MC subcomponents be elicited and reliably measured via a battery of MC tests?
>
> RQ2: To what extent do identifiable factors underlie scores from the MC Test Battery, and which aspects of L2 MC might they represent?

METHOD

Participants

For the main study, 112 L1 Mandarin speakers of L2 English (*M* age = 22.9; SD = 2.6) completed our MC Test Battery (see section 'Materials: MC Test Battery development'). Most (99/112) were postgraduates enrolled for or undertaking study at a total of eight UK universities. The remainder were undergraduates studying at these universities. All except seven were studying social science degrees. A further 16 participants started the study but later dropped out.

The MC Test Battery was also completed by 31 L1 English speakers (*M* age = 39.7, SD = 16.3) who had learned English as their home and main language from birth. Data from these participants were used for establishing scoring parameters for two receptive tests involving acceptability judgement and for corroborating the researcher-expected responses in all multiple-choice receptive tests. Most L1 English participants (25/31) were currently in or recently retired from full-time employment, while six were postgraduate students at a UK university studying for PhDs in Education, History, and an MSc in Global Marketing. They were recruited as a convenience sample and were all British citizens and L1 English speakers living in various parts of the UK. Although these participants were, on average, older than the L2 participants, their range of ages, professions, and localities within the UK is likely to have yielded a minimally idiosyncratic picture of contemporary L1 'British' English metaphor use.

Materials: MC Test Battery development

Construct selection and representativeness: First, all of Low's (1988) and Littlemore and Low's (2006a) metaphor-related skills/sub-competences were listed (by study heading and sub-heading) and evaluated for testability. Tests were then developed if all of the following criteria were met: (i) the construct could be meaningfully operationalized in language; (ii) metaphor was central (rather than peripheral) to the construct; (iii) the construct did not involve controversial language or ideas; and/or (iv) the construct did not likely generate obvious overlap with a test for another construct. The eventual battery contained tests measuring nine of the authors’ general constructs, comprising six out of Low’s 10 metaphor-related skills and operationalizing three of Littlemore and Low’s four sub-competences. The full list of MC constructs tested or omitted, with full rationales, is provided in the Supplementary Material (see also O’Reilly 2017). A summary of example constructs that we decided to omit, and the reasons for this is given in Table 1.

The final battery contained untimed receptive and productive tests, in the written modality (i.e. reading and writing). As such, the tests allowed participants maximum opportunity to access their linguistic resources and reduced potential anxiety related to time pressure or speaking. The

Table 1: Examples of constructs from Low (1988) and Littlemore and Low (2006a) rejected for test development

Issue	Example	Decision
(1) *Construct could not be meaningfully operationalized in language*		
	Demonstratives (Littlemore and Low 2006a: 157–58)—we were unable to develop, for example, multiple-choice options to pinpoint a single correct metaphorical meaning and plausible distractors to distinguish metaphorical uses of ‘this/that’ based on their literal senses of physical closeness/distance (e.g. in ‘the dress has got *this/that* awful pattern on it’ both ‘this’ and ‘that’ seem to metaphorically convey distaste).	Test not developed
(2) *Metaphor peripheral (rather than central) to the construct*		
	Hedges (Low 1988: 133)—a test was piloted but removed to reduce the size of the battery, since it was the least centrally concerned with metaphor, and due to the difficulty of evaluating and scoring responses. As Low noted, the notion of a correct/incorrect use of ‘literally’ is problematized by the fact this hedge may signify either literal or non-literal meanings.	Test not developed
(3) *Construct involved controversial language/ideas*		
	Socially sensitive metaphors (Low 1988: 133–34), register (e.g. insults, swearing) (Littlemore and Low 2006a: 102–09) —we heeded Littlemore and Low’s (2006a) warning that ‘it is the brave teacher who says . . . “here is what you may encounter” and “here is how to get your own back”’ (p. 105) and avoided exposing participants to such forms.	Test not developed
(4) *Construct had obvious overlap with another*		
	Conventional exploitation of Vehicle term features covered both by Low (1988: 130–31), and to some extent Littlemore and Low (2006a: 109–10), under ‘naturalness’.	One test developed (Test 3-Vehicle acceptability-R)

written modality was also conceptually and methodologically more straightforward for construct operationalization, as it afforded control over the language elicited and reduced confounding factors that spoken or listening performance can add. Although it is acknowledged that, ideally, L2 MC in the spoken modality should also be elicited, this would have placed unreasonable demands on participants and was beyond the scope of the current study.

We used the same lexicon across receptive and productive tests in order to reduce conflation of vocabulary knowledge with performance across modes. To this end, two versions of the MC Test Battery were created that counterbalanced items across receptive and productive tests: metaphors used as receptive test items in MC Test Battery version 1 (completed by half the participants) were used as productive items in version 2 (completed by the other half), and vice-versa. Participants were randomly assigned one of these two versions, which were statistically equivalent (see section 'Analysis') and the same in all other respects.

Table 2 presents an overview of the MC Test Battery that the participants undertook. This includes 'test name' tagged as receptive '-R' or productive '-P'; 'construct tested', and its 'operationalization'; 'part'[3]; '*k*' items per construct measured; 'skill type' receptive or productive; 'question type', for receptive tests: multiple-choice gap-fill, explain-the-meaning, and acceptability judgement; for productive tests: limited production gap-fill (Bachman and Palmer 1996: 54), also known variously as constrained, constructed, or controlled production; and, in the final column, information about how items were 'scored' (0–1 or 0–2 points). For all multiple-choice questions, four response options were used (Lee and Winke 2013), as these require less reading time than five options, reduce the chances of correct random guessing compared with three options, and our pilot participants reported preferring four options.

In the remainder of this subsection, we provide three illustrative examples of test items. For the full (long) and refined (shorter) MC Test Battery (versions 1 and 2), and record of items retained after data cleaning, we refer readers to www.iris-dababase.org (Marsden *et al.* 2016), while O'Reilly (2017) contains a 30-page account of the operationalization of tests, including the selection and refinement of items and scoring decisions.

Test example (1). Test 1-Phrasal verbs-R and -P were developed to operationalize test-takers' ability to recognize and recall (metaphorical) phrasal verb particles (Littlemore and Low 2006a: 162–66). Five phrasal verb particles frequently associated with conceptual metaphors ('up', 'off', 'out', 'in', 'down') were chosen and, using Gardner and Davies' (2007: 358–59) list of Frequency and Coverage of Top 100 Phrasal Verb Lemmas in BNC, four phrasal verb forms for each particle were selected from different frequency bands, for wider coverage. Sentences were then developed in which these verbs were used metaphorically, confirmed by applying MIPVU (Steen *et al.* 2010). For the receptive test, the multiple-choice format was used, and distractors evoking related concepts were selected:

Table 2: MC Test Battery overview

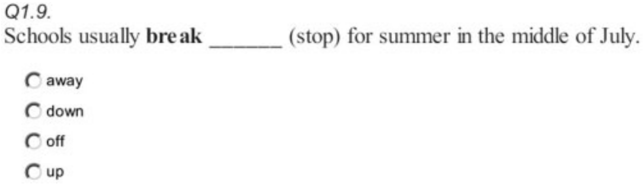
Test name	Construct(s) tested	Operationalized as test of ability to:	Part	*k*	Skill type	Question type	Scored
Test 1-Phrasal verbs-R	Grammatical competence—phrasal verbs (Littlemore and Low 2006a: 162–66)	Recognize metaphorical phrasal verb particles	A	10	Receptive	Multiple-choice (gap-fill)	0, 1
Test 1-Phrasal verbs-P		Recall metaphorical phrasal verb particles	B	10	Productive	Limited production (gap-fill)	
Test 2-Metaphor layering-R	Awareness of multiple layering in metaphors (Low 1988: 134) Ability to construct plausible meanings (Low 1988: 129)	Understand the meaning of linguistic metaphors	A(a)	6	Receptive	Limited production (explain-the-meaning)	0, 1
		Recognize the most relevant aspect of meaning for understanding metaphors	A(b)	6	Receptive	Multiple-choice (gap-fill)	
		Recognize endings to garden path sentences (fig-lit)	B	6	Receptive	Multiple-choice (gap-fill)	
		Recognize endings to garden path sentences (fig-fig)	C[a]	6	Receptive	Multiple-choice (gap-fill)	
Test 3-Vehicle acceptability-R	(Knowledge of the boundaries of conventional metaphor [including] . . .) Knowledge of which features of the vehicle Y can be exploited conventionally and which cannot (Low 1988: 130–31)	Rate the acceptability of semantic exploitations of vehicles	A[b]	16	Receptive	Rating scale (acceptability judgement)	0, 1
	Knowledge of Vehicle acceptability across different word classes (Low 1988: 131)	Rate the acceptability of Vehicles across different word classes	B[b]	12	Receptive	Rating scale (acceptability judgement)	
Test 4-Topic/ Vehicle-R	Awareness of acceptable Topic and Vehicle combinations (Low 1988: 132)	Rate the acceptability of Vehicles as analogies for given Topics	A	6	Receptive	Rating scale (acceptability judgement)	0, 1
Test 4-Topic/ Vehicle-P		Produce Vehicles as analogies for a given Topics	B	6	Productive	Limited production (gap-fill)	0, 1, 2

(*Continued*)

Table 2: *(continued)*

Test name	Construct(s) tested	Operationalized as test of ability to:	Part	*k*	Skill type	Question type	Scored
Test 5-Topic transition-R	Textual competence: marking the edges of a text-figurative language in topic transition (Littlemore and Low 2006a: 144–49)	Recognize idioms/proverbs/sayings in topic transition	A	6	Receptive	Multiple-choice (gap-fill)	0, 1
Test 5-Topic transition-P		Produce idioms/proverbs/sayings in topic transition	B	6	Productive	Limited production (gap-fill)	0, 1, 2
Test 6-Heuristic-R	Illocutionary (heuristic) functions (Littlemore and Low 2006a: 126–29)	Recognize similes used to perform heuristic functions	A	6	Receptive	Multiple-choice (gap-fill)	0, 1
Test 6-Heuristic-P		Produce similes to perform heuristic functions	B	6	Productive	Limited production (gap-fill)	0, 1, 2
Test 7-Feelings-R	Illocutionary (ideational) functions (Littlemore and Low 2006a: 112–16)	Recognize metaphors that convey feelings about information	A	6	Receptive	Multiple-choice (gap-fill)	0, 1
Test 7-Feelings-P		Produce metaphors that convey feelings about information	B	6	Productive	Limited production (gap-fill)	0, 1, 2
Test 8-Idiom extension-R	Illocutionary (imaginative) functions (Littlemore and Low 2006a: 129–32)	Recognize extensions of the literal senses of idioms	A	6	Receptive	Multiple-choice (gap-fill)	0, 1
Test 8-Idiom extension-P		Produce extensions of the literal senses of idioms	B	6	Productive	Limited production (gap-fill)	0, 1, 2
Test 9-Metaphor continuation-R	Interactive awareness of metaphor (Low 1988: 134–35)	Recognize continuations of metaphor in discourse	A	6	Receptive	Multiple-choice (gap-fill)	0, 1
Test 9-Metaphor continuation-P		Produce continuations of metaphor in discourse	B	6	Productive	Limited production (gap-fill)	0, 1, 2

[a] Seen by test-takers as part of ‘Section 2 Part B’.
[b] Seen by test-takers as ‘Section 3’ (i.e. not divided into Parts A and B).



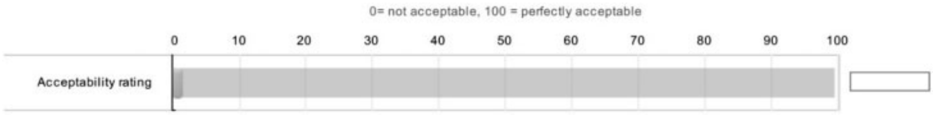
Q1.9.
Schools usually **break** _______ (stop) for summer in the middle of July.

○ away
○ down
○ off
○ up

Here, test-takers might be distracted by 'away' if they reasoned along the lines of departure or exit, and 'down' or 'off' if they focused on concepts of inactivity and deactivation (Kövecses and Szabó 1996).

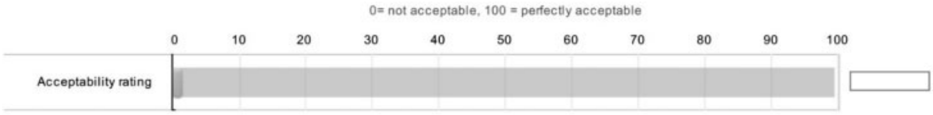
Test example (2): While most receptive tests used a multiple-choice format, some (tests 3 and 4) used acceptability judgement. For example, the first 16 items of Test 3-Vehicle acceptability-R operationalized the ability to rate the acceptability of different exploitations of the Vehicle (Low 1988: 130–31). Test-takers were instructed that 'English native speakers often use expressions which mix ideas and concepts in . . . a strange way'. They were then given two examples with explanations of higher and lower acceptability expressions. Then, for the actual test items, they were requested to '. . . rate the acceptability of [each] expression in bold by dragging the slide. An acceptable expression is one that an English native speaker might use in the context of the sentence'. For example (Part A):

Q3.2
He **slipped** into a depression.



0= not acceptable, 100 = perfectly acceptable

Q3.12.
Her hair had almost **arrived at** being grey.

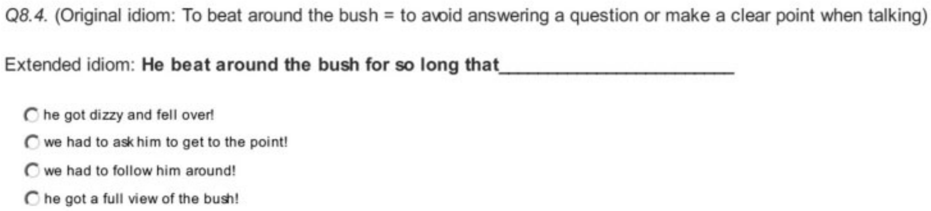


0= not acceptable, 100 = perfectly acceptable

All items in this test were linguistic metaphors sourced (and adapted) from the British National Corpus (Davies 2004) that link to various conceptual metaphors. For the items above, the conceptual metaphor was CHANGE IS MOTION (O'Reilly 2017).

The higher/lower acceptability of items was corroborated by the 31 L1 participants' (L1ers') ratings, which were also used to establish scoring parameters. For example, on average the L1ers rated Q3.2 as 87.65 per cent acceptable (SD = 25.68) and Q3.12 as 8.16 per cent acceptable (SD = 13.14). The variability of these L1 ratings was also considered during test development, whereby items with high variability (with SD > 25 per cent, see section 'Analysis') were eliminated. Thus, Q3.2 had relatively high variation in L1 judgements and so was not retained in the final battery. Q3.12, on the other hand, was retained, and to score '1' (*correct*), a test-taker (L1 or L2) needed to rate the item 0–21 per cent acceptable (L1ers' $M \pm 1 \times \text{SD}$). Although the somewhat decontextualized nature of items may be a limitation, following the process described above, we at least ensured we were dealing with the most clear-cut cases (from the L1ers' perspective) of 'acceptable' and 'unacceptable' vehicle extensions.

Test example (3): Test 8-Idiom extension -R and -P were designed to operationalize learners' ability to playfully extend the literal sense of idioms, part of illocutionary (imaginative) functions (Littlemore and Low 2006a: 129–32). After an introduction explaining that speakers often play with fixed phrases, an example showing how the *literal* sense of an idiom can be extended, with an explanation, test-takers were asked to choose the 'best' option for extending a series of six idioms (-Receptive/-R, see below), and to provide their own extensions for a further six (-Productive/-P). All idioms sourced were Macmillan English Dictionary phrases, evidence of their current usage:



Q8.4. (Original idiom: To beat around the bush = to avoid answering a question or make a clear point when talking)

Extended idiom: **He beat around the bush for so long that______________________**

○ he got dizzy and fell over!

○ we had to ask him to get to the point!

○ we had to follow him around!

○ he got a full view of the bush!

For receptive items such as Q8.4, the 'best' answer was the most appropriate extension of the literal sense (here, the first option). We developed at least one distractor involving the figurative sense (the second option) and at least one less acceptable literal extension (the third and fourth options). 'Best' answers were corroborated by responses from the L1ers, while distractors were evaluated and refined using the L2 pilot participants' responses. For productive items, test-takers supplied their own extension (see below for scoring information).

In sum, the test and study design were driven by the need to target Low/Littlemore's constructs and to build on the limitations of previous measures (e.g. using an L1 homogenous group of L2 learners, using a MIP to verify item metaphoricity).

Piloting the tests: Items and tests were refined through three (pre-)pilot studies involving L2 English speakers (L1 Mandarin) and L1 English speakers: (i) three L2ers completed draft test versions; (ii) their responses were discussed by the first author and two L1ers, and between authors; (iii) two different L1ers completed revised drafts. In a subsequent pilot study, 10 L2ers and four L1ers completed a revised draft, with three of the L2ers and two L1ers also thinking aloud (see O'Reilly (2017) for details).

Scoring: All receptive questions were scored ‘1’ (*correct*) or ‘0’ (*incorrect*). ‘Correct’ answers to receptive questions were corroborated as such by L1ers’ responses in the pilot studies, with the exception of responses to Test 2-Metaphor layering-R part A(a) ‘explain the meaning’ questions, which were scored via different raters (see below).

Limited production responses were scored by three raters according to the extent to which they fulfilled the task as either ‘2’ (*correct*), ‘1’ (*partially correct*), or ‘0’ (*incorrect*). L1 English speaker raters were chosen partly for convenience and partly to reduce the chances of introducing additional variance (such as cross-linguistic and cross-cultural influences) into the corroboration of scoring for receptive and productive tests (i.e. fulfilling a similar function to the L1 participants). We emphasize here that we do not intend to prescribe L1 judgements, but rather, to use them as a point of comparison. The three raters scored responses to a total of 78 items comprising (i) 72 items in productive tests (not including 20 items in Test 1-Phrasal verbs-P, 10 each for versions 1 and 2, which elicited pre-specified prepositions scored *correct*/*incorrect*, and so did not require rater corroboration); and (ii) six explain-the-meaning items in Test 2-Metaphor layering-R part A(a), a receptive test.

Raters were instructed to focus on evaluating the meaning of responses and told that grammatical accuracy and spelling were not part of the scoring criteria. In total, three inter-rater reliability estimates and one intra-rater reliability estimate were calculated according to the procedure described immediately below (for all estimates, see RQ1: Eliciting and reliably measuring subcomponents of L2 MC). Rater 1 was the first author (L1 English). Rater 2 (L1 English) was trained by scoring several tests using the scoring criteria and a glossary of key terms (O’Reilly 2017). Raters 1 and 2 discussed their decisions to promote consistency. Rater 2 then scored all tests (versions 1 and 2); her scores were compared with rater 1’s decisions to calculate the first inter-rater reliability estimate and identify disagreements. For each disagreement, raters 1 and 2 then reconsidered their original decision, working independently, without the other rater’s original score. The revised decisions were then compared to calculate a second inter-rater reliability estimate, and remaining disagreements resolved during face-to-face meetings to arrive at final rater 1 and 2 decisions. Five months later, rater 3 (L1 English) was trained and tasked to score all responses. A third inter-rater reliability estimate was calculated by comparing rater 3’s decisions with rater 1 and 2’s final decisions (see section ‘Participants’). Five months after that, rater 1 conducted a second pass, scoring all limited production responses again. An intra-rater reliability estimate was calculated as the agreement between rater 1 and 2’s final decisions and rater 1’s second pass.

Procedure

Main data collection took place from June to November 2015. All L2 participants were offered a small reward (£5 cash or Amazon voucher), informed that the tests were unconnected to their studies, and invited to receive feedback on test answers after the data collection period. Tests were administered via Qualtrics. In order to maximize participation and minimize anxiety, L2 participants were offered the choice of completing tests in pre-arranged lab sessions or at home. Thirty-five L2ers completed the test in the lab and 77 at home, with test setting having no detectable effect on the data, $F(2, 16) = 1.03$, $p = 0.415$ (Munzel and Brunner's 2000 robust MANOVA).[4] The total time needed to complete the MC Test Battery in the lab was 1.5–2 hours (testing sessions were a similar duration in Littlemore 2001) and home-based participants were requested to complete the MC tests in one sitting. Items within each test were ordered from easy-to-difficult according to the pilot data.

Analysis

Data were analysed using R programming language (R Core Team 2019) with 13 packages, one additional script (see Supplementary Material), and Microsoft Excel.

Data cleaning. In order to address RQ1 (on eliciting and reliably measuring subcomponents of L2 MC), and maximize the validity of MC tests as measures of the proposed constructs, the MC Test Battery data were first cleaned in six stages (Table 3).

Table 3: Summary of six stages of data cleaning

Data cleaning stage	Criteria for removal	Removals	
		Type	Deleted/total
(1) Items with rating scale outliers (t3r only)	L1 speakers' acceptability ratings SD > 25%	Item	10/28
(2) Participant outliers	Individual score < 3 × SD for that test	Participant	1/112
(3) Item analysis	Diff 0.33–0.67 *with* disc <0.30; L2 diff > L1 diff	Item	14/226
(4) Low L1 speakers' scores	Problematically low average L1 test score	Test	1/16
(5) Instrument reliability	Inconsistent items (up to ordinal $\omega = 0.74$ or 4 items)	Item Item Item Item	33/84 (v1-R) 29/83 (v2-R) 11/38 (v1-P) 9/37 (v2-P)
(6) Distractor analysis	Multiple-choice items with no distractor < 0	Item	0/139

Notes. t3r = Test 3-Vehicle acceptability-R; diff = item difficulty (ranging from 0 to 1); disc = item discriminability (ranging from −1 to +1); v1 and v2 = test versions 1 and 2, respectively; -R and -P = receptive and productive.

Stage 1: 10 receptive items with wide variation in L1 acceptability ratings were deleted.

Stage 2: One L2 participant with an extremely low score for Test 2-Metaphor layering-R was removed.

Stage 3: 14 items (from Test 1-Phrasal verbs-R and -P, Test 2-Metaphor layering-R, Test 4-Topic/Vehicle-R and -P, and Test 6-Heuristic-P) that were poor discriminators between high and low scoring L2 test-takers[5] (Aiken 2003), or for which the L2ers scored more highly than the L1ers, were removed.

Stage 4: For Test 4-Topic/Vehicle-P, the L1 participants produced many literal descriptions (e.g. ‘CCTV cameras are the security of the building’) rather than analogies as intended (e.g. ‘eyes’), indicating a construct validity problem, and so this test was not included in the final battery (see O’Reilly (2017) for further details).

Stage 5: Ordinal omega (ω) was chosen as the instrument reliability index (McNeish 2018) since tests elicited ordinal item response data, items varied in their relationship with the overall test construct (congeneric scales), and overall test scores were calculated from equally-weighted items. Items were removed if they lowered a test’s reliability, stopping when ordinal ω exceeded 0.74, a recently proposed general (not absolute) threshold for instrument reliability in L2 research (Plonsky and Derrick 2016). If removing such items reduced the number of items in a test to below four, items were not removed.

Stage 6: All remaining 68 multiple-choice items had at least one well-performing distractor, luring more lower than higher ability test-takers, so there were no further removals.

After these six stages of data cleaning, the final MC tests comprised data from 4 to 18 items per test version (see section ‘RQ1: Eliciting and reliably measuring subcomponents of L2 MC’).

Calculating overall percentage scores and test version parity. After data cleaning, participants’ scores were converted to percentages out of the total marks available per test. Also, four separate overall totals were calculated: MC-Receptive (versions 1 and 2) and MC-Productive (versions 1 and 2).

Since the two different MC test versions measured the same constructs, used the same, counterbalanced set of lexical exemplars, and showed no statistically significant differences, $F(2,\ 13) = 1.16$, $p = 0.321$ (Munzel and Brunner’s 2000 robust MANOVA),[6] they had sufficient conceptual and statistical parity to be merged to provide one receptive score and one productive score.

Exploratory factor analysis. In order to address RQ2 (on factors underlying the MC Test Battery scores), we conducted an Exploratory Factor Analysis (EFA) on the L2 response data (only). Exploratory, rather than confirmatory, factor analysis was used since no previous L2 MC study provides a predetermined basis on which to establish hypotheses for a Confirmatory Factor Analysis. Factor loadings of greater than 0.32 were interpreted (Tabachnick and Fidell 2013). Principal Axis Factoring, which entails no distributional assumptions (Fabrigar *et al.* 1999), was used to deal with uni- and multivariate nonnormality. As recommended by various authors (Loewen and Gonulal 2015; Plonsky and Gonulal 2015), multiple criteria were used to determine how many factors to retain. While parallel analysis (Horn 1965) and the scree plot suggested retaining one factor only, Joliffe's (1972) eigenvalues-greater-than-0.7 rule, suggested as many as nine. In the end, we used Kaiser's greater-than-one rule (Kaiser 1960), suggesting four factors, because this offered the most interpretable and theoretically informative solution. Since tests measured aspects of human cognition and were expected to be correlated (Plonsky and Gonulal 2015), the solution was rotated using Direct Oblimin.

RESULTS

RQ1: eliciting and reliably measuring subcomponents of L2 MC

Instrument reliability The reliability of the overall MC Test Battery receptive and productive items was high (version 1 MC-R ordinal omega/$\omega = 0.85$, MC-P ordinal $\omega = 0.79$; version 2 MC-R ordinal $\omega = 0.85$, MC-P ordinal $\omega = 0.87$).[7] The consistency of item sets for individual tests was more modest on average (but still high), with some variation (ordinal $\omega = 0.56$–0.92, $M = 0.72$, SD $= 0.10$, Mdn $= 0.71$, IQR $= 0.17$, see Supplementary Material and www.iris-database.org (see Marsden *et al.* (2016) for full list).[8]

Rater reliability checks for production tests 5-9 and 2 A(a) questions 'Substantial' agreement (Landis and Koch 1977) was observed for rater 1 and 2's 3,075 'initial decisions' (weighted kappa/$Kw = 0.63$)[9] and rater 3's 3,075 decisions versus rater 1 and 2's 'final scores' ($Kw = 0.65$). Intra-rater reliability, measured as rater 1's 'second pass' versus rater 1 and 2's 'final scores' ($Kw = 0.81$), rounded to Landis and Koch's (1977) lower bound of 0.81 for 'almost perfect' agreement. Agreement between rater 1 and 2's 838 'revised decisions' was 'moderate' ($Kw = 0.47$), meaning some disagreements persisted until the final discussion stage (see section 'Materials: MC Test Battery development').

Descriptive statistics. Table 4 provides for each test and the total test battery: participant numbers, item numbers, average scores, measures of spread, and rank difficulties.

Figure 1 shows a bee swarm plot of mean test scores, with plotted points spreading widthways for participants achieving the same percentage score. Means are shown by the large circles and the bars indicate one standard deviation above and below the mean. Figure 2 shows median test scores, with bars extending from the lower to upper quartiles.

Table 4: MC Test Battery descriptive statistics after data cleaning and reliability checks, $N = 112$

	K	Mean		Median		Spread	
Test	Total (v1, v2)	%	Rank	%	Rank	SD	IQR
T1-Phrasal verbs-R	10 (4, 6)	58.4	5	58.3	6	24.8	25
T1-Phrasal verbs-P	10 (6, 4)	44.3	10	50	7	24.4	41.7
T2-Metaphor layering-R[a,b,c]	10 (10, 10)	55.9	6	50	7	20.8	30
T3-Vehicle acceptability-R[a,b]	18 (18, 18)	38.5	13	38.9	12	18.1	29.2
T4-Topic/Vehicle-R	8 (4, 4)	60.0	4	50	7	27.2	25
T5-Topic transition-R	8 (4, 4)	69.9	1	75	1	26.9	50
T5-Topic transition-P	8 (4, 4)	40.4	12	37.5	13	28.7	50
T6-Heuristic-R	9 (4, 5)	64.7	2	75	1	27.2	30
T6-Heuristic-P	8 (4, 4)	55.5	7	62.5	4	25.8	37.5
T7-Feelings-R	8 (4, 4)	63.2	3	75	1	27.5	25
T7-Feelings-P	8 (4, 4)	47.4	9	43.8	11	25.3	37.5
T8-Idiom extension-R	10 (5, 5)	28.6	15	20	14	26.5	40
T8-Idiom extension-P	12 (6, 6)	31.9	14	16.7	15	32.5	58.3
T9-Metaphor continuation-R	10 (5, 5)	51.8	8	60	5	25.6	20
T9-Metaphor continuation-P	9 (5, 4)	41.7	11	45	10	27.9	50
MC Test Battery-R[a,b]	91 (58, 61)	54.7	1	55.4	1	12.3	17.2
MC Test Battery-P[a]	55 (29, 26)	43.5	2	43.3	2	17.6	25.1

[a] Normally distributed, Shapiro–Wilk test > 0.01.
[b] Versions 1 and 2 contained the same T2 and T3 items.
[c] N = 111 (version 1, n = 55; version 2, n = 56).

Table 4 and Figures 1 and 2 reveal that, statistically, the easiest tests were Test 5-Topic transition-R, Test 6-Heuristic-R, and Test 7-Feelings-R, whereas the most difficult were Test 8-Idiom extension-R and -P, the latter of which also had the most widely dispersed scores. The overall battery scores were descriptively higher for receptive than productive.

RQ2: factors underlying the MC Test Battery scores

Figure 3 shows the pattern matrix loading for the four-factor solution, which explained 34 per cent of the total variance in test scores.

The model was adequate by various indicators (e.g. Comparative Fit Index = 1.058) and five (out of 12) loadings are likely to have been statistically significant according to Stevens' (2002) rule-of-thumb (loadings > 0.512 when $N > 100$). We took a principled approach to factor interpretation and compared information about the loading strength and descriptions of what each variable (test within the factor) aimed to measure, as well as the lower 95 per

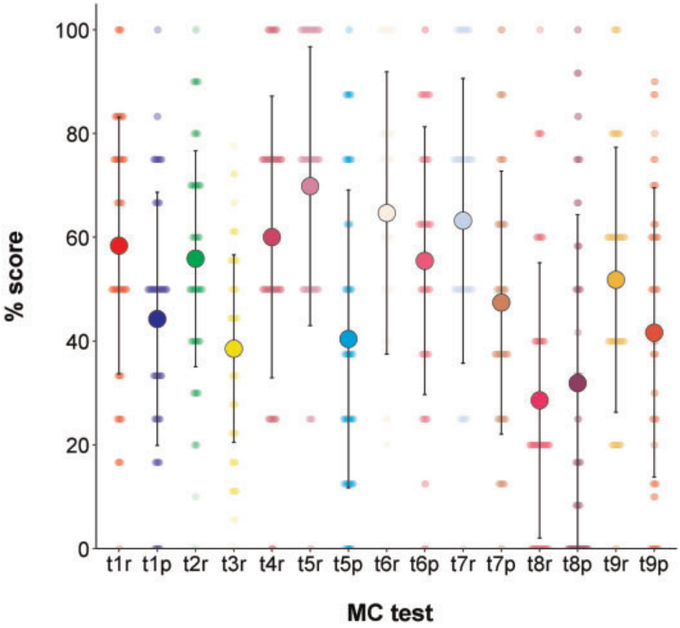


Figure 1: MC test scores (means)

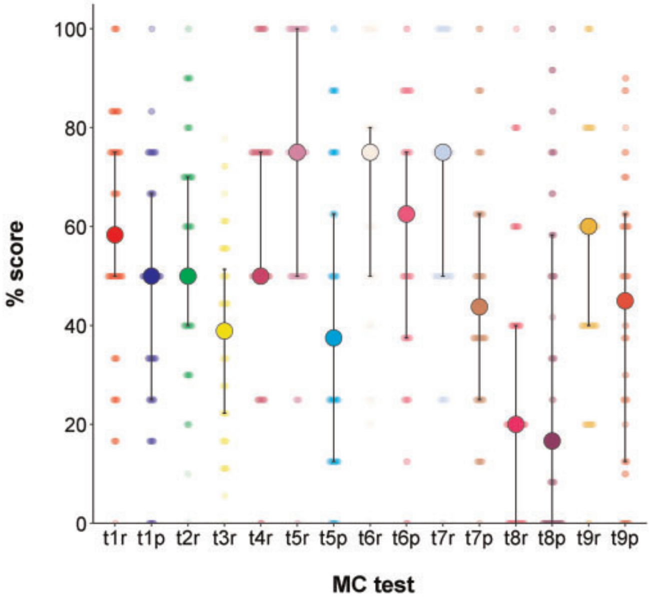


Figure 2: MC test scores (medians)

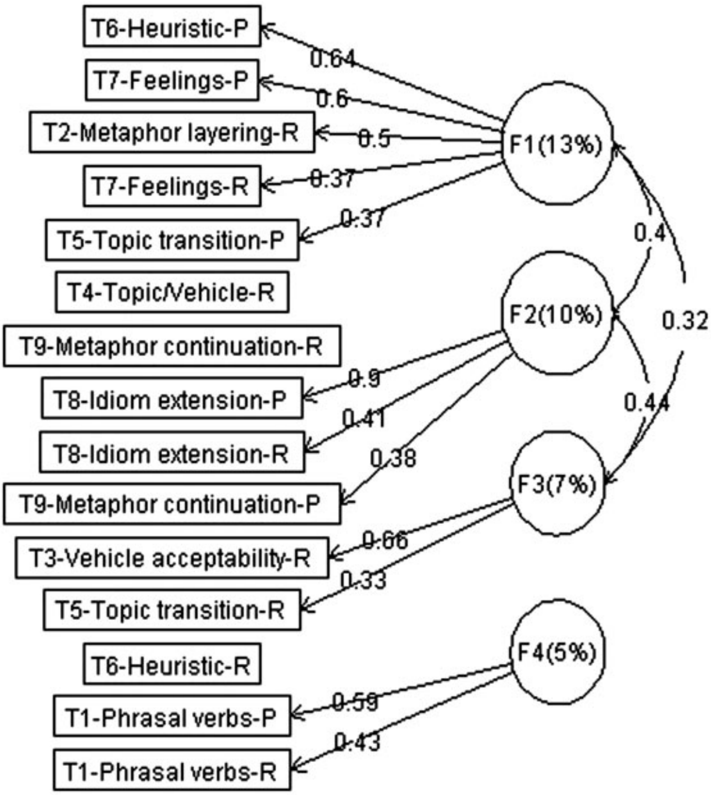


Figure 3: Diagram of MC factors/loadings, Direct Oblimin rotated (34 per cent total variance explained)

cent confidence intervals of 5,000 bootstrap resamples as a measure of internal replicability (Zientel and Thompson 2007).

Factor 1 was named 'Productive Illocutionary MC', since its two strongest markers concerned the production of linguistic metaphors for illocutionary purposes, namely teaching young children about biological, natural, and physical entities, and conveying feelings about people, situations, and experiences. This factor had high item-within-factor consistency (test version 1 ordinal omega/$\omega = 0.75$, 26 items; test version 2 ordinal $\omega = 0.81$, 26 items). Factor 2 was labelled 'Metaphor Language Play', on account of the ludic aspects to all three tests loading on this factor. The consistency of items-within-factor was high (version 1 ordinal $\omega = 0.88$, 16 items; version 2 ordinal $\omega = 0.85$, 15 items). Factor 3 was named 'Topic/Vehicle Acceptability' because Topics, Vehicles, and acceptability judgements connected its two loading variables. This factor had high item consistency (version 1 ordinal $\omega = 0.78$, 22 items; version 2 ordinal $\omega = 0.81$, 22 items). Finally, factor 4 was labelled 'Grammatical MC', given that loading tests measured knowledge of relatively fixed forms, (metaphorical) phrasal verbs. This factor had relatively low-to-mid item consistency (version 1 ordinal $\omega = 0.55$, 10 items; version 2 ordinal $\omega = 0.60$, 10 items), with fewer contributing items than the other factors.

DISCUSSION

This study elicited and measured L2 MC, making a novel contribution to the plethora of studies that have investigated other aspects of L2 competence and knowledge, such as morphosyntax, the lexicon, pragmatics, and pronunciation. MC test development is equally important as these other agendas, if we wish to improve our understanding of the nature of MC, our capacity to investigate MC-related questions in the future, and our ability to measure MC in, for example, instructed contexts.

The study is, to the best of our knowledge, the first attempt to design a battery of tests to elicit theoretically motivated model of L2 MC drawing on Low/Littlemore's metaphor-related skills/sub-competences. It is also, to our knowledge, the first use of EFA to uncover latent subcomponents of L2 MC, providing insight into the nature of the L2 MC construct. Rigorous data cleaning and reliability testing provided the most valid and reliable set of items we were able to attain. On the basis of data from these items, the EFA suggested that L2 MC construct was underpinned by four latent variables: Productive Illocutionary MC, Metaphor Language Play, Topic/Vehicle Acceptability, and Grammatical MC.

We first discuss the elicitation and measurability of the MC construct (RQ1), examining details about test development, reliability, and analysis. Such details are important for future use or adaptation of these tests and also the development of not only MC tests, but of L2 tests more generally. We then consider the latent factors of MC that emerged from our tests (RQ2), providing insight into the nature of the MC construct as theorized by Low/Littlemore.

Eliciting and reliably measuring subcomponents of L2 MC (RQ1)

In many ways, we improved on past attempts to measure MC: the group of participants was large and homogenous in terms of all having the same L1. The use of an elicited approach inevitably means that the researchers need to make decisions about scoring criteria. This was the case in our study, but we took several steps to minimize the impact of our own subjectivity (31 L1 participants used to corroborate our proposed 'correct' answers to receptive tasks, two additional raters used to maximize consistency of scoring productive responses). Our various scoring, rater, reliability, and data cleaning procedures allowed us to remove problematic items and data; and we tested a wider range of components of MC than previous studies.

However, we acknowledge that the scope of the MC Test Battery was still constrained, and we could not cover all of Low/Littlemore's constructs (see section ‘Materials: MC Test Battery development’). One example was the absence of a test of ‘ability to interpret and control “hedges” [e.g., genuinely, sort of, literally]’ (Low 1988: 133), which we considered peripheral to core MC, since such tuning devices frequently occur with non-metaphorical language. Similarly, spelling and grammatical accuracy of productive metaphor were not accounted for in our approach to scoring, even though they featured in Low/Littlemore’s accounts of MC, particularly grammatical accuracy. Nevertheless, the MC Test Battery still achieved considerable coverage of the authors’ constructs, operationalizing six out of 10 of Low’s metaphor-related skills and three out of four of Littlemore and Low’s subcompetences.

The large size of our initial test battery (*K* items = 236, comprising 92 different and 52 shared version 1 and 2 items) enabled us to remove problematic test items without compromising the breadth of construct coverage too severely. Our systematic and staged data cleaning resulted in the removal of 38 per cent of the items seen by test-takers, which aligns with Littlemore’s reductions (of 29, 33, and 50 per cent) after piloting, and highlights steps needed to maximize the validity and reliability of tests measuring MC constructs.

The internal consistency of the total test battery was higher than many of the MC tests in the 33 studies we reviewed prior to the study. (However, we note that direct comparison between our findings and previous findings is complicated by the mix of participants (L1 and L2, proficiencies) and different coefficients used across the sample of studies we reviewed.)

The reliability of many of our *individual* MC tests was, however, somewhat lower than that of the total battery and more similar to several previous single instruments that we reviewed. This could perhaps be due to the smaller number of items in these tests and/or their focus on (overly) specific MC subcompetences. The variation in reliability that we observed between individual tests aligns with Littlemore’s (2001) L2 English MC tests ($\alpha = 0.31$–0.90 with 5–25 items per test) and may be explained in some, though not all, instances. For example, the notably high instrument reliability of Test 8-Idiom extension-R and -P is probably due to the fact that for creative metaphor tasks such as these, scores are frequently skewed towards the lower end of scales (Beaty and Silvia 2013). Indeed, our learners were fairly consistent in having low scores across all Test 8-Idiom extension-R and -P items, yielding comparatively high inter-item correlations.

Of critical interest for the wider field of applied linguistics, and particularly all subdomains that involve test construction including effectiveness-of-intervention studies, is our use of ordinal omega as our instrument reliability coefficient. This choice allowed us to account for the non-continuous nature of item responses and variation in how strongly items relate to the construct (i.e. violation of tau equivalence). This increased the accuracy of our estimates. It also increased the estimates themselves by an important margin. While our data did not meet the conditions necessary for use of Cronbach alpha, had we not checked this and inappropriately used this coefficient, MC test reliability would have been underestimated by an average of over 30 per cent (Mdn ordinal $\omega = 0.71$ vs. Mdn $\alpha = 0.50$).

This is a clear illustration for the field of applied linguistics that the choice of reliability coefficient is critical. We chose to use ordinal omega because we did not have any prior empirical models from which to posit minor dimensions when estimating reliability; instead, our approach was to assume unidimensionality and use ordinal omega to estimate MC test and factor reliability. While practical, one may raise a conceptual objection to this assumption, since most items were eventually found to relate to *both* MC test and factor dimensions. If we had *not* assumed unidimensionality, a suitable coefficient would have been (ordinal) omega hierarchical, which 'attempts to parse out the variability attributable to subfactors and calculate reliability for a general factor that applies to all items' (McNeish 2018: 417). Had we used the hierarchical variation of ordinal omega, reliability estimates would have been only slightly lower for MC tests (3–4 per cent), and the factors Productive Illocutionary MC (3 per cent), Metaphor Language Play (3 per cent), and Topic/Vehicle Acceptability (6 per cent), but more substantially lower for Grammatical MC (15 per cent), suggesting that minor dimensions may have impacted reliability estimates most strongly for this latter factor. While a comprehensive comparison of approaches to reliability estimation is beyond the scope of the current study, we provide the full list of estimates by various coefficients and information about residuals in the Supplementary Material.

We also sought to conduct extensive rater reliability checking, where multiple raters scored metaphor productions. Decisions were largely consistent, reflecting clear scoring criteria. However, rater 1 and 2's revised decisions after initial disagreements were less consistent, with many discrepancies persisting until the final discussion stage, where they were eventually resolved. While there may be many reasons why raters vary in scoring MC productions for quality, task fulfilment, etc., for now we simply note that such variation can exist, and is likely for both elicited and naturalistic production. This should be explored if the field wishes to converge on robust and efficient measurement of metaphor production.

We used L1 speakers' responses for verifying researcher-intended 'best' answers and for identifying problematic items and tests (data cleaning stages 1, 3, and 4). Such an approach is not common, but was used similarly by Chen and Lai (2015). In Test 3-Vehicle acceptability-R (Low 1988: 130–31), participants rated the acceptability of linguistic metaphors, some of which extended Vehicles to intentionally convey semantically odd aspects of Vehicle terms (e.g. Q3.10 'The theory was the colour of brick') or presented Vehicles in an unconventional word class (e.g. Q3.17 'He freshened his ideas'). Here, we used L1 speakers' ratings as our benchmark for establishing a range of acceptability within which the L2 participants should rate in order to receive a score of '1' (see section 'Materials: MC Test Battery development'), and retained only the most clear-cut highly un/acceptable items. While this worked well for most items, there was considerable variation in the L1 speakers’ reported acceptability of four phrases that are, nevertheless, attested in the British National Corpus (Davies 2004).

One possible reason for the variability here is that despite the careful phrasing of the items and rating criteria, refined through think-aloud protocols during piloting (see section ‘Materials: MC Test Battery development’), the rubric needs further improvement to ensure participants have a sufficiently similar sense of appropriacy or acceptability in mind throughout, a key challenge when contextual information is limited, in this case, to a word/phrase embedded within a sentence. Since discourse context will likely shape the notion of acceptability regardless of whether an interaction involves exclusively L1 or L2 speakers, mixed L1/L2, Lingua Franca or any other variety of English, future MC test developers might further specify the test instructions by providing information about imagined interlocutors, their relationships, etc. (e.g. *Imagine you are speaking in English, in this café in Amsterdam [picture], at 12:30pm on January 15th, with two close friends, Anna—a 21 year old, female, L1 German speaker of low proficiency, and Hiroki—an 18-year old, male, L1 Japanese speaker of intermediate proficiency*). Test-takers might also be informed about scoring criteria (e.g. *To gain 1 point, your rating must fall within the mean plus or minus one standard deviation of the ratings from a group of 30 adult L2 English raters with a mix of L1 backgrounds*). While these steps may help learners know which broad types of ‘norms’ they should align with, this approach may never be practically achievable within an elicited methodology, given the endless list of contextual details that might shape the acceptability of a particular metaphor, although we tentatively suggest a possible way forward below (see section ‘Limitations and future directions’).

Our data indicate that variation associated with acceptability ratings seems to be more pronounced with certain metaphors and creative extensions of metaphor, and less so with others, where there is more homogeneity within the variety of English we used to establish scoring criteria. Nevertheless, the variability observed could lead us to question the validity of Vehicle Acceptability as a generalizable or even ‘real’ construct, since reliable measurement at *group* level could be unfeasible, an issue that Low recognized early on, ‘a language learner who simply transferred the word-class preferences of his or her first language might well be seen as either consciously innovating (possibly for humorous purposes), or else as making an error in the second language’ (1988: 131). However, it may be that the extension of Vehicles, semantically and across word classes, varies between individuals, or perhaps between varieties of English, even more than Low (1988) surmised. Clearly, considerable time has elapsed since Low’s article and our finding is perhaps reflected in more recent perspectives on language use and notions of acceptability that question whether these are in fact definable, and/or whether they exist with real homogeneity at a group level. Even with empirical, corpus-based evidence that specific linguistic metaphors are common, and a theoretically homogenous group of adult L1 speakers from the same geographical locality, we found acceptability of Vehicle extensions to be a highly subjective sub-competence to elicit and measure. Given their importance, we consider these issues in more depth elsewhere (O'Reilly and Marsden 2020b).

We finish this section by revisiting the question of the role of a 'target' with regard to metaphor and MC. For pedagogical purposes, we suggest that there can certainly be target linguistic metaphors (e.g. metaphorical phrasal verbs, codifiable and attested collocations and idiomatic expressions, discourse domain-specific stock phrases) and clearly defined functions of metaphor (e.g. explaining abstract concepts to children, playing with metaphor for humour, invoking idioms/proverbs/sayings to facilitate a change of topic) that teachers and learners might wish to work with. Crucially, if communicative success (rather than a constrained set of norms or pre-fabricated language) is the strongest determinant of what should be learned and taught (Kathpalia and Carmel 2011) then for most learners this means learning to use metaphor successfully in multi- (rather than mono)lingual settings. Here, one complete set of linguistic norms cannot determine communicative success, and the ability to express oneself with metaphor, as opposed to using particular, 'fixed' metaphors (Hall 2014), is key.

This notion of MC manifests in some of the dimensions identified in the current study, for example, with productive metaphor language play involving humour (see section 'Factors underlying the MC Test Battery scores (RQ2)'), for which speakers/writers must have knowledge of and engage with the specific discourse context and interlocutors' cultures, personalities, level of English, etc. In such cases, there is no 'one-size-fits-all', ready-made set of linguistic metaphors.

With tests of receptive knowledge (and with Test 1-Phrasal verbs-P also), construct operationalization necessitated that we select specific, 'target' metaphors and standards by which to measure comprehension of these, expressed in terms of linguistic criteria (test-takers must demonstrate comprehension of a specific meaning of a linguistic metaphor) rather than more functional criteria (where various interpretations could be acceptable). Although we measured receptive knowledge in relation to elicited L1 norms, and productive knowledge using several L1 judges (although somewhat flexibly, since task completion and communication were emphasized), we stress that the pedagogy surrounding the tests we developed might be much more open to variety. Teachers and learners using our tests would be free to adapt and explore, for example, functionally equivalent ELF variants of the metaphors we utilized, and variability due to discourse context, language background, speakers' dialect(s), and any other factors of interest.

The points we have set out here align with recent conceptualizations of viewing multilingualism from within a transdisciplinary framework (Group 2016), and chime with Low's (1988) conceptualization of competence, not as mastery of a particular set of metaphors, but as the extent to which a speaker (L1, L2, etc.) is accepted by the social groups they interact with and within, and/or rejected 'on account of excruciating boringness' (Low 1988: 129). We encourage future MC test developers seeking to elicit and evaluate metaphor more authentically to explore MC in these terms. Yet, where linguistic criteria apply, our key message is to consider which 'norms' are appropriate for informing measurement, whether these be L1, L2, ELF, Bajan English, Geordie, or any other.

Factors underlying the MC Test Battery scores (RQ2)

The EFA showed four latent factors in the 15 receptive and productive MC measures, which we interpreted as Productive Illocutionary MC, Metaphor Language Play, Topic/Vehicle Acceptability and Grammatical MC. While the 34 per cent total variance explained by these four factors was smaller than average compared within SLA research more widely (58 per cent in Plonsky and Gonulal 2015), our factors were conceptually coherent and the solution adequate by all post-hoc criteria suggesting a robust model that can be used to shed some light on the nature of L2 MC.

Creative and conventional metaphor dimensions that had been observed in L1 MC factor structures (Beaty and Silvia 2013) were broadly evident in our factor solution (although direct comparison is confounded by different participants (L1 vs. L2), instrumentation, and study features). The most creative factor we uncovered was Metaphor Language Play, defined largely by Test 8-Idiom extension-P, which involved producing appropriate and, typically, humorous extensions of the literal senses of idioms. This required participants to adapt existing knowledge to discourse situations for which they had few pre-existing linguistic solutions, a skill that characterizes general Communicative Language Ability (Bachman 1990), suggesting Metaphor Language Play may draw on a wide range of communicative sub-competences (Canale and Swain 1980). In contrast, the most conventional metaphor dimension emerging from our data was Grammatical MC, since its markers elicited knowledge of fixed, dictionary-codified forms. Only one factor, Topic/ Vehicle Acceptability, had a loading test that used acceptability judgement (Test 3-Vehicle acceptability-R), and so the issues with this format (described above on pages 28–29) pertain exclusively to this factor.

In general, our factors lend some support to Low's and Littlemore and Low's seminal accounts of L2 MC. Three of our four factors were indeed marked by receptive and productive tests of three unique constructs described by Littlemore and Low (2006a): Productive Illocutionary MC—marked by both productive and (to a lesser extent) receptive heuristic tests (pp. 126–29); Metaphor Language Play—by receptive and productive idiom extension tests (pp. 129–32); and Grammatical MC—by receptive and productive phrasal verbs tests (pp. 162–66).

However, apart from Grammatical MC, the other three factors also had loadings from tests developed to measure metaphor in relation to other language competence components discussed by the authors, suggesting additional conceptual links between several of the proposed constructs. Specifically, we found that both 'awareness of multiple layering in metaphors' (Low 1988:

134) and 'producing (figurative) topic transitions/textual competence' (Littlemore and Low 2006a: 144–49) were associated with Productive Illocutionary MC; that 'productive interactive awareness of metaphor' (Low 1988: 134–35) was associated with Metaphor Language Play; and that both 'knowledge of the boundaries of conventional metaphor' (Low 1988: 130–31) and 'recognizing (figurative) topic transitions/textual competence' (Littlemore and Low 2006a: 144–49) were associated with Topic/Vehicle Acceptability.

In terms of balance between receptive and productive skills represented within our factors, there was just one factor (Topic/Vehicle Acceptability) that was loaded exclusively by receptive tests. This may have been influenced to some extent by the fact that we could not develop a productive test corresponding to two of the tests in the battery (Test 2-Metaphor layering-R; Test 3-Vehicle acceptability-R). Nonetheless, Test 5-Topic transition-R loaded on Topic/Vehicle Acceptability while its productive equivalent did not, but rather, was more strongly associated with Productive Illocutionary MC. This could suggest that *recognizing* versus *creating* idioms/proverbs/sayings in topic transition may be distinct constructs.

Despite these insights, we strongly emphasize that our factors did not explain approximately two-thirds of the test data, suggesting possible explanatory roles for other (non-MC) L2 knowledge aspects, such as breadth and depth of vocabulary knowledge and/or general proficiency. We address this question in O'Reilly and Marsden (2020a).

Limitations and future directions

The scope of our study was limited in several dimensions. We covered only a subset (albeit most) of Low/Littlemore's skills and sub-competences, and while the MC Test Battery elicited competence with metaphor, simile, and idiom, tropes such as metonymy were not a focus. Future research might broaden the battery's scope more generally, and also extend it to the oral modality and other L1 groups to explore potential cross-linguistic influences.

A perplexing, though not unusual, methodological challenge that we faced was the range (of between one and nine) different factor solutions that was offered by the different factor retention criteria that we considered in order to select which model to interpret, as advised by statistics guidance (Field *et al.* 2012; Loewen and Gonulal 2015; Plonsky and Gonulal 2015). In response, we advocate further research here, and invite researchers who have conducted EFAs, in any domain, to make their data available for re-analysis, as we have done on www.iris-database.org (Marsden *et al.* 2016). Our use of the bootstrapping technique suggested some level of reproducibility of our analysis. However, the real proof of the theoretical pudding will come via external, rather than internal replication of the solution.

A final limitation concerns the effectiveness of acceptability judgement as an item format in MC testing. As discussed above, while an efficient format compared with multiple-choice and gap-fill production, acceptability judgement brought certain problems. First, L1 judgements varied greatly on several items designed to be acceptable/unacceptable; these items could not then be used in evaluating L2 learners' sensitivity to acceptable extensions of Vehicle terms. Secondly, even though the instructions stated how test-takers should approach the notion of acceptability, there may always exist some conceivable discourse context in which a seemingly unacceptable metaphor is acceptable, and vice-versa. Thirdly, our use of British, adult, L1 speakers restricted our notion of correctness to one particular variety (or collection of varieties) of English.

Future studies might address these limitations via a more tailored approach using combined methodologies (albeit at the possible expense of the sample size). For example, the researcher might first collect metaphors that learners and interlocutors produce in naturalistic speech (e.g. Bell 2005), and then via elicitation techniques (e.g. stimulated recall) have learners rate the acceptability of the metaphors produced in these interactions, where information about the discourse context, interlocutor etc. is known.

However, in the context of the test battery as a whole, the use of acceptability rating in two tests is not likely to have had a large impact on the validity of the MC dimensions that we identified. This is because Test 3-Vehicle acceptability-R loaded substantially on one factor only, its items comprising only 12 per cent (18/146) of those retained from the battery, while Test 4-Topic/Vehicle-R, with items comprising 5 per cent (8/146) of those retained, did not load substantially on any factor. Nevertheless, we acknowledge that in spite of the benefits of the acceptability judgement format for eliciting metaphor comprehension (e.g. allowing for greater item coverage than multiple-choice), its limitations in this study (e.g. test-takers possibly having different discourse contexts in mind when rating) were evident.

CONCLUSION

The main achievement of the current study is a refined MC Test Battery measuring several of Low/Littlemore's metaphor-related skills/sub-competences, with items and tests indexed for reliability, difficulty, and discriminability. The instrument (in whole or part) can now be used by researchers and teachers wishing to tap MC. The refined battery (available on www.iris-database.org, Marsden *et al.* 2016), which survived data cleaning, had an overall high level of instrument reliability and a robust scoring protocol. This may be particularly useful as it is shorter than the full, initial battery.

In developing and refining the MC Test Battery, we followed a systematic methodology involving several item, test, and participant outlier analyses. Our study is also likely to be one of the first in the field of applied linguistics to estimate instrument reliability using a superior alternative to Cronbach's alpha, as recommended in the wider psychometric testing literature. Together, these steps allowed for identification of underperforming or misfit items and participants, and for their subsequent removal to maximize validity and reliability of the tests.

We encourage future researchers wishing to elicit and measure L2 MC, or other constructs, to undertake and report on the various techniques used here, despite the high number of item and data deletions that are likely to result. Refining our toolkit in such ways is a sign of the maturity of the field. The development of robust instrumentation is essential for conceptualizing and identifying constructs and, as such, plays an important role in theory-building. It also serves the more applied need to elicit and measure core components of language competence, as reliably as we know how.

NOTES

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

2 We considered Kathpalia and Carmel's (2011) essay question an MC 'instrument', since it was devised specifically to elicit metaphor, unlike similar studies involving language corpora (Nacey 2013).

3 The MC Test Battery appeared as one long 'test' with 'sections' (MC tests) and 'parts' (e.g. receptive/productive questions).

4 Specifically, the non-significant test statistic meant that we could not reject the null hypothesis (in this case, H_0 = Completing MC tests in 'lab' and 'home' settings has no effect on scores), which was sufficient for our purposes, but not equivalent to strong evidence for the null hypothesis of the kind offered by Bayesian approaches and/or external replication (Wetzels *et al.* 2009; see Morgan-Short *et al.* (2018) for an example).

5 'Poor' discriminators were items that satisfied two conditions: (i) a low discriminability score (<0.30) and (ii) a mid-range difficulty index (0.33–0.67). Item discriminability and difficulty correlate such that the former decreases as the latter approaches both 1 and 0 (i.e. an item for which all or no test-takers gain marks does not discriminate). Thus, we tolerated low discriminability scores for items outside the mid-range (see Aiken (2003) for further details).

6 Final items retained for Test 1-Phrasal verbs-P and Test 9-Metaphor continuation-P (versions 1 and 2) were chosen to ensure statistical parity (see Note 4), but were not the most internally consistent.

7 A further six version 1 and six version 2 MC-R items that were negatively correlated with the overall scale were deleted before the reliability estimates could be specified.

8 These statistics include one ordinal Cronbach's alpha estimate of 0.62 for Test 6-Heuristic-R (version 2), since the scaleStructure function in the UserFriendlyScience package (see Supplementary Material) failed to identify a model while computing ordinal omega for this test.

9 Weighted kappa (*Kw*) with linear weights was used since decisions involved ordinal data with scoring disagreements penalized equally throughout the scale (e.g. '0' vs. '1' and '1' vs. '2').

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

Conflict of interest statement. None declared.

REFERENCES

Aiken, L. R. 2003. *Psychological Testing and Assessment*. Allyn and Bacon.

Aleshtar, M. H. and **H. Dowlatabadi**. 2014. 'Metaphoric competence and language proficiency in the same boat,' *Procedia - Social and Behavioral Sciences* 98: 1895–904.

Azuma, M. 2005. *Metaphorical Competence in an EFL Context*. Toshindo.

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing*. Oxford University Press.

Bachman, L. F. and **A. Palmer**. 1996. *Language Testing in Practice*. Oxford University Press.

Beaty, R. E. and **P. J. Silvia**. 2013. 'Metaphorically speaking: Cognitive abilities and the production of figurative language,' *Memory & Cognition* 41: 255–67.

Bell, N. 2005. 'Exploring L2 language play as an aid to SLL: A case study of humor in NS-NNS interaction,' *Applied Linguistics* 26: 192–218.

Boers, F., **J. Eyckmans**, and **H. Stengers**. 2007. 'Presenting figurative idioms with a touch of etymology: More than mere mnemonics?,' *Language Teaching Research* 11: 43–62.

Canale, M. and **M. Swain**. 1980. 'Theoretical bases of communicative approaches to second language teaching and testing,' *Applied Linguistics* 1: 1–47.

Chen, Y.-C. and **H. L. Lai**. 2015. 'Developing EFL learners' metaphoric competence through cognitive-oriented methods,' *International Review of Applied Linguistics in Language Teaching* 53: 415–38.

Cronbach, L. J. 1951. 'Coefficient alpha and the internal structure of tests,' *Psychometrika* 16: 297–334.

Davies, M. 2004. *British National Corpus.* Oxford University Press.

Deignan, A., **E. Semino**, and **S.-A. Paul**. 2019. 'Metaphors of climate science in three genres: Research articles, educational texts, and secondary school student talk,' *Applied Linguistics* 40: 379–403.

Fabrigar, L. R., **D. T. Wegener**, **R. C. MacCallum**, and **E. J. Strahan**. 1999. 'Evaluating the use of exploratory factor analysis in psychological research,' *Psychological Method*. 4: 272–99.

Field, A., **J. Miles**, and **Z. Field**. 2012. *Discovering Statistics Using R*. Sage.

Gardner, D. and **M. Davies**. 2007. 'Pointing out frequent phrasal verbs: A corpus-based analysis,' *TESOL Quarterly* 41: 339–59.

Gibbs, R. W. 2014. 'Why do some people dislike conceptual metaphor theory?,' *Cognitive Semiotics* 5: 14–36.

Hall, C. J. 2014. 'Moving beyond accuracy: From tests of English to tests of "Englishing",' *ELT Journal* 68: 376–85.

Horn, J. L. 1965. 'A rationale and test for the number of factors in factor analysis,' *Psychometrika* 30: 179–85.

Joliffe, I. T. 1972. 'Discarding variables in a principle component analysis. I: Artificial data,' *Applied Statistics* 21: 160–73.

Kaiser, H. F. 1960. 'The application of electronic computers to factor analysis,' *Educational and Psychological Measurement* 20: 141–51.

Kathpalia, S. S. and **H. L. H. Carmel**. 2011. 'Metaphorical competence in ESL student writing,' *RELC Journal* 42: 273–90.

Kövecses, Z. 2010. *Metaphor: A Practical Introduction*, 2nd edn. Oxford University Press.

Kövecses, Z. and **P. Szabó**. 1996. 'Idioms: A view from cognitive semantics,' *Applied Linguistics* 17: 326–55.

Lakoff, G. and **M. Johnson**. 1980. *Metaphors We Live by*. University of Chicago Press.

Landis, J. R. and **G. G. Koch**. 1977. 'The measurement of observer agreement for categorical data,' *Biometrics* 33: 159–74.

Lee, H. and **P. Winke**. 2013. 'The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test,' *Language Testing* 30: 99–123.

Littlemore, J. 2001. 'Metaphoric competence: A language learning strength of students with a holistic cognitive style?,' *TESOL Quarterly* 35: 459–91.

Littlemore, J. and **G. D. Low**. 2006a. *Figurative Thinking and Foreign Language Learning*. Palgrave Macmillan.

Littlemore, J. and **G. D. Low**. 2006b. 'Metaphoric competence, second language learning, and communicative language ability,' *Applied Linguistics* 27: 268–94.

Loewen, S. and **T. Gonulal**. 2015. 'Exploratory factor analysis and principle components analysis' in L. Plonsky (ed.): *Advancing Quantitative Methods in Second Language Research*. Routledge, pp. 182–212.

Loewen, S., **S. Li**, **F. Fei**, **A. Thompson**, **K. Nakatsukasa**, **S. Ahn**, and **X. Chen**. 2009. 'Second language learners' beliefs about grammar instruction and error correction,' *The Modern Language Journal* 93: 91–104.

Low, G. D. 1988. 'On teaching metaphor,' *Applied Linguistics* 9: 125–47.

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 and C. Juchem-Grundmann (eds): *Applied Cognitive Linguistics in Second Language Learning and Teaching. AILA Review*, vol. 23. John Benjamins, pp. 155–73. https://www.worldcat.org/title/applied-cognitive-linguistics-in-second-language-learning-and-teaching/oclc/664668490.

Macmillan English Dictionary. 2019. Available at https://www.macmillandictionary.com/. Accessed: 30 November 2019.

Marsden, E., **A. Mackey**, and **L. Plonsky**. 2016. 'The IRIS Repository: Advancing research practice and methodology' in A. Mackey and E. Marsden (eds): *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages*. Routledge, pp. 1–21.

McNeish, D. 2018. 'Thanks coefficient alpha, we'll take it from here,' *Psychological Methods* 23: 412–33.

Morgan-Short, K., **E. Marsden**, **J. Heil**, **B. I. Issa II**, **R. P. Leow**, **A. Mikhaylova**, **S. Mikołajczak**, **N. Moreno**, **R. Slabakova**, and **P. Szudarski**. 2018. 'Multisite replication in Second Language Acquisition research: Attention to form during listening and reading comprehension,' *Language Learning* 68: 392–437.

Munzel, U. and **E. Brunner**. 2000. 'Nonparametric tests in the unbalanced multivariate one-way design,' *Biometrical Journal* 42: 837–54.

Nacey, S. 2013. *Metaphors in Learner English*. John Benjamins.

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

O'Reilly, D. and **E. Marsden**. 2020a. 'Metaphoric competence in a second language: A construct explained by vocabulary knowledge and general proficiency?' (manuscript under review).

O'Reilly, D. and **E. Marsden**. 2020b. 'The role of L1 data in L2 metaphoric competence assessment' (manuscript under review).

Pitzl, M.-L. 2016. 'World Englishes and creative idioms in English as a lingua franca,' *World Englishes* 35: 293–309.

Plonsky, L. and **D. Derrick**. 2016. 'A meta-analysis of reliability coefficients in second language research,' *The Modern Language Journal* 100: 538–53.

Plonsky, L. and **T. Gonulal**. 2015. 'Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis,' *Language Learning* 65: 9–36.

Plonsky, L., **E. Marsden**, **D. Crowther**, **S. Gass**, and **P. Spinner**. 2019. 'A methodological synthesis and meta-analysis of judgment tasks in second language research,' *Second Language Research* 1–39. https://journals.sagepub.com/doi/pdf/10.1177/0267658319828413.

Pollio, H. R. and **B. Burns**. 1977. 'The anomaly of anomaly,' *Journal of Psycholinguistic Research* 6: 247–60.

Pollio, H. R. and **M. Smith**. 1979. 'Sense and nonsense in thinking about anomaly and metaphor research,' *Bulletin of the Psychonomic Society* 13: 323–6.

Pollio, H. R. and **M. Smith**. 1980. 'Metaphoric competence and complex human problem solving' in R. P. Honeck and R. R. Hoffman (eds): *Cognition and Figurative Language*. Lawrence Erlbaum, pp. 365–92.

Pragglejaz. 2007. 'MIP: A method for identifying metaphorically used words in discourse,' *Metaphor and Symbol* 22: 1–39.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available at https://www.R-project.org/. Accessed: 30 November 2019.

Raykov, T. and **G. A. Marcoulides**. 2019. 'Thanks coefficient alpha, we still need you!,' *Educational and Psychological Measurement* 79: 200–10.

Rodriquez, M. C. and **Y. Maeda**. 2006. 'Meta-analysis of coefficient alpha,' *Psychological Methods* 11: 306–22.

Saito, K. and **L. Plonsky**. 2019. 'Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis,' *Language Learning* 69: 652–708.

Semino, E., **Z. Demjén**, **A. Hardie**, **S. A. Payne**, and **P. E. Rayson**. 2018. *Metaphor, Cancer and the End of Life: A Corpus-Based Study*. Routledge.

Steen, G., **A. G. Dorst**, **J. Berinke Herrmann**, **A. A. Kaal**, **T. Krennmayr**, and **T. Pasma**. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.

Stevens, J. P. 2002. *Applied Multivariate Statistics for the Social Sciences*, 4th edn. Erlbaum.

Tabachnick, B. G. and **L. Fidell**. 2013. *Using Multivariate Statistics*, 6th edn. Pearson Education.

The Douglas Fir Group. 2016. 'A transdisciplinary framework for SLA in a multilingual world,' *The Modern Language Journal* 100: 19–47.

Thomas, M. 2006. 'Research synthesis and historiography: The case of assessment of second language proficiency' in J. M. Norris and L. Ortega (eds): *Synthesizing Research on Language Learning and Teaching*. John Benjamins, pp. 279–300.

Wetzels, R., **J. G. W. Raaijmakers**, **E. Jakab**, and **E-J. Wagenmakers**. 2009. 'How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test', *Psychometric Bulletin & Review* 16: 752–60.

Wheeler, D. L., **M. Vassar**, **J. A. Worley**, and **L. L. B. Barnes**. 2011. 'A reliability generalization meta-analysis of coefficient alpha for the Maslach Burnout Inventory,' *Educational and Psychological Measurement* 71: 231–44.

Zhao, Q., **L. Yu**, and **Y. Yang**. 2014. 'Correlation between receptive metaphoric competence and reading proficiency,' *English Language Teaching* 7: 168–81.

Zientel, L. R. and **B. Thompson**. 2007. 'Applying the bootstrap to the multivariate case: Bootstrap component/factor analysis,' *Behavior Research Methods* 39: 318–25.