Applied Linguistics, 2023, XX, 1–22
https://doi.org/10.1093/applin/amad067
Advance access publication 23 November 2023
Article

Playing with second language metaphor: An exploration with advanced Chinese learners of English

David O'Reilly[1,*] and Luling Yan[2]

[1]Centre for Advanced Studies in Language and Education, Department of Education, University of York, York YO10 5DD, UK
[2]Kensington Elementary School, Waxhaw, NC, USA

*E-mail: david.oreilly@york.ac.uk

The present study continues research that takes non-serious language more seriously (Cekaite and Aronsson 2005) by focusing on a central second language (L2) Metaphoric Competence factor, Metaphor Language Play (MLP). For willing learners, MLP offers a diversity of benefits (Bushnell 2009; Bell 2012a) despite being one of the most challenging Metaphoric Competence aspects (O'Reilly and Marsden 2021). While studies provide rich descriptions of naturally occurring MLP, elicitation approaches are needed to target comprehension/production of specific forms/meanings/usages and types of play, for example, comprehension of US sitcom humour (Dore 2015). With 69 advanced first-language Mandarin L2 English learners, we addressed these issues via an Exploratory Factor Analysis to uncover hitherto unknown relationships between written/spoken/receptive/productive MLP measures, and a thematic analysis of the linguistic, conceptual, and metalinguistic themes in learners' MLP. The findings revealed three underlying MLP factors, two positively related, and a rich set of linguistic, conceptual, and metalinguistic themes. The implications of findings for future research and pedagogy are discussed.

Introduction

Language Play (hereafter LP), somewhat paradoxically, is a normal abnormality, a standard part of language characterized by deviation and manipulation of formal and contextual 'norms' (Crystal 1996, 1998; Cook 2000; Belz 2002). While play *with* language involves word/sound/meaning manipulation, for example, 'spill the *pasta*' in a BBC report on an Italian mafioso's secret disclosure (Naciscione 2020), play *in* (a) language involves language use to engage in playful games, narratives, joking, and so on (Bell 2012b). LP can concern linguistics, semantics, and/or pragmatics (Cook 2000; Sullivan 2000), emerging for enjoyment/fun (ludic) or rehearsal/skill development/form learning (Lantolf 1997; Borner and Tarone 2001). As a sign of healthy existence and development (Crystal 1996; Cook 1997), numerous benefits await the second language (L2) learner willing to seriously engage in what is often regarded as a non-serious, irrelevant means of self-expression (Cekaite and Aronsson 2005).

LP is rich with metaphor and other figurative language, instances where X is treated as if it were in some ways Y (Low 1988). These include extensions of the literal sense of idioms (e.g. 'I've been sitting on the fence so long my bottom is beginning to hurt', Littlemore and Low 2006: 130), metaphors continued throughout an interaction (Low 1988), and other phenomena not typically codified in regular or specialist dictionaries but prevalent in interactions, media, popular music, fiction, political interviews, and other discourse domains.

In recent work operationalizing theoretically motivated descriptions of L2 Metaphoric Competence via a battery of written-mode tests, Metaphor Language Play (hereafter MLP and (M)LP to denote LP-focused studies involving metaphor) emerged as one of four latent sub-components of the wider competence for first-language (L1) Mandarin L2 English learners (O'Reilly and Marsden 2021). In a further study, MLP had a strong, positive relationship with vocabulary depth (O'Reilly and Marsden 2023). However, the generalizability of learners' MLP abilities from written- to spoken-mode tasks, the linguistic and conceptual nature of learners' playful metaphor comprehensions/productions, and signs of metalinguistic awareness were not investigated. It is these facets of the MLP dimension from O'Reilly and Marsden (2021, 2023) that we further explore in the present study, taking a deep dive into the latent relationships between written/spoken/receptive/productive MLP for advanced L1 Mandarin L2 English learners and their characteristic linguistic/conceptual/metalinguistic themes.

We begin by reviewing literature on the MLP construct, its benefits for L2 learning, elicitation and naturalistic approaches, and on MLP patterns in an example of popular media, used to devise one of the current study's tests. After summarizing the research gaps, questions to be addressed and methods, we report the results with an integrated discussion, outlining future research directions and tentative pedagogical implications. Finally, limitations and concluding remarks are provided.

Literature review

The Metaphor Language Play (MLP) construct

MLP concerns the ability to play with conventional and novel metaphor in language (Werkmann Horvat *et al.* 2022) and to creatively innovate, for example 'you are the weakest link! [gameshow slogan]' to tease a boyfriend about his university group project (Bell 2005: 202–3). (M)LP studies contain qualitative analyses of naturalistic learner-driven play and/or quantitative analyses of elicited comprehension/production of researcher-chosen play via tasks/tests.

Research using naturalistic approaches has documented the role of (M)LP in learners' growing, personalized, multicompetence development with already-known forms (Belz 2009); their sociolinguistic competence development, for example, via experimentation with different voices (Bushnell 2009); and instances of increasingly creative play and meta-reflection, for example, a 'clone' metaphor used by friends giving rise to a mini-vocabulary lesson (Bell 2005: 209–12). In English as a Lingua Franca research especially, corpus-based examples show speakers drawing from multilingual resource pools and interlocutor knowledge to perform situation-/speaker-specific MLP (Pitzl 2016).

Research using elicitation approaches targets play with specific metaphors/figurative language. Littlemore and Low (2006), for example, identified re-literalizing for comic effect as a key language play opportunity, but via a small-scale study found that advanced L2 English learners had limited success in adapting Prodromou's (2003) manipulated idioms to pre-specified contexts (e.g. adapting 'bring home the bacon' to a person earning a lot for their family). More recently, mixed-proficiency (Common European Framework of Reference/CEFR A2-C2) L1 Mandarin L2 English learners had some success in comprehending/producing operationalizations of MLP as described by Littlemore/Low, but this remained a difficult competence area (O'Reilly 2017; O'Reilly and Marsden 2021).

Elicitation studies, also, show rich MLP development. When invoking an idiom/saying/proverb to close a topic, several of O'Reilly's (2017) learners misapplied plural/third person/possessive-s

such as '[when in Rome] do as Romes do/as Rome does/do in Romes as Rome does', indicating constituent/non-holistic word processing (Conklin and Schmitt 2012). While sometimes disruptive (Miller 2011), L1–L2 formulaic language differences can in fact be equally communicatively successful. For example, when re-literalizing 'the ball is in my court, the problem is …', O'Reilly's (2017) L1ers preferred racquet sports whereas the L2ers opted for throwing sports, football, and baseball; with 'go and break a leg, in fact …' the L1ers favoured the alliterative collocation 'break both', the L2ers 'break two'.

As an important sub-component of elicited Metaphoric Competence, MLP seems to reside on the more creative end of the spectrum, positively related to latent Productive Illocutionary Metaphoric Competence and Topic-Vehicle Acceptability factors but not Grammatical Metaphoric Competence (O'Reilly and Marsden 2021). Recent research (O'Reilly and Marsden 2023) has revealed a strong link between MLP and Word Associates Test vocabulary depth (Read 1998), suggesting both draw on a common associative thinking ability (Carroll 1993; Littlemore 2001, 2008; Littlemore and Low 2006). However, since these findings pertain to written-mode MLP, the nature of the MLP construct across written-/spoken-modes, and for higher- rather than mixed-proficiency learners is unknown.

The benefits of engaging in MLP

Engaging in MLP is especially worthwhile for L2 learners. Bushnell (2009), drawing on Tarone (2000), argues that LP, in general, lowers affective SLA barriers (e.g. anxiety), aids memorability by creating strong traces/triggering associations (Craik and Lockhart 1972), affords opportunities for experimenting with target 'voices' (Bakhtin 1981), promotes interlanguage system destabilization and restructuring, allows for safely committing face threatening acts (see also Bell 2012b), and for collaborative, form-focused attention.

Bell (2012a) points to evidence of bizarre materials (McDaniel *et al.* 1995) and humorous images/language (Schmidt and Williams 2001; Strick *et al.* 2010) aiding L2 form-meaning acquisition, her own study found that 16 mixed-proficiency, adult, L2 English learners had better post hoc elicited recall of playful rather than serious language involved in incidental reflections over eight weeks. Other studies show that L2 learners were better at deciphering comic strip/pun/riddle meanings when first made aware of potential double meanings (Lucas 2005; Tocalli-Beller and Swain 2007).

Usefully, MLP seems to make target language *more* memorable, and non-target/mislearned language *less* memorable and easier to unlearn. This is because both LP (Tarone 2005) and metaphor (Low 1988; Littlemore and Low 2006; Macarthur 2010) direct overt attention to form-meaning-usage links, destabilize interlanguage, and prevent fossilization. However, since manipulated language formulas are socially riskier and more processing-intensive than unamended variants (e.g. formulaic sequences), learning benefits will likely appear in the longer rather than shorter term (Bell 2012b).

In computer-mediated communication, (M)LP affords specific opportunities for acronym and emoticon use, and for additional time to decipher and produce playful language compared (Bell 2012b). In O'Reilly and Marsden's (2021) Metaphoric Competence Test Battery, an MLP-loading test involved producing metaphors to continue humorously cryptic, online (written-mode) chat (Low 1988), for example, jokingly comparing a colleague to a wizard and discussing a friend's pregnancy in code.

MLP in sitcom humour

L2 learners encounter much MLP in film, television, popular music, fiction, social media (e.g. memes), and journalism. US/UK situation comedies (sitcoms), for example, contain many jokes in which a figurative expression is (mis)understood literally. This phenomenon was analysed in Dore's (2015) innovative work, which combined Conceptual Metaphor Theory (CMT), Blending Theory (BT), and the General Theory of Verbal Humour (GTVH) to show how the scriptwriters/actors/directors of the US sitcom *Friends* playfully exploit metaphor for comedic effect and to reinforce the six main characters' traits.

CMT (Lakoff and Johnson 1980) posits unidirectional source-to-target domain mappings to explain linguistic metaphors, BT supposes dynamic mental spaces in which entrenched associations are activated and modified with the unfolding thought and discourse (Fauconnier and Turner 1996, 1998, 2002), while GTVH (Attardo and Raskin 1991; Attado 1994, 2001) stipulates that a Script Opposition parameter reveals incongruity in humour, which is resolved by a Logical Mechanism, only partly in the case of metaphor. Through their combined use, Dore (2015) provided a comprehensive analysis of the complexity of the metaphors, meaning shifts, and blended imagery within the dynamic nature of conversational humour, robust to the shortcomings of a single theory.

The findings revealed the 'funny and sometimes grotesque' (Dore 2015: 202) nature of the humour to establish character idiosyncrasies (e.g. Joey's simple-mindedness and food/sex interests) and contextually insensitive behaviour (Low 1988). As Joey compares dating to sampling ice cream flavours, a recently divorced Ross adds a jab line, 'I don't know if I'm hungry or horny', blending literal-figurative imagery, further exploited in Chandler's crude punchline, 'stay out of my freezer!'. Although such humour relies upon the entrenched, Western cliché of men objectifying women (e.g. as food), analytically speaking, Joey's metaphor is highly creative, exploiting a blend of figurative/literal meanings and involving a lack of semantic resolution triggering humour.

While little is known about how L2ers might notice, comprehend, and acquire such scripted/acted MLP, research suggests that L1 speakers more easily notice and remember bizarre, emotionally laden than mundane language in soap-operas/sitcoms (Bates *et al.* 1980, as cited in Bell 2012a). Vocabulary knowledge undoubtedly plays a crucial role. While O'Reilly and Marsden's (2023) learners' vocabulary knowledge would, theoretically, well-equip them for MLP comprehension in films and television (cf. Webb and Rodgers 2009a, 2009b), this was not measured, and it is unclear how well written-mode MLP comprehension would generalize to the spoken-mode, an empirical question that the current study addresses.

Summary and research questions

Research has shown that MLP is a central but challenging part of L2 Metaphoric Competence. For willing learners, MLP offers a diversity of benefits alongside inevitable challenges. While studies using naturalistic approaches have provided rich MLP descriptions, by design, they do not reveal the extent of learners' abilities with different forms/meanings/types of play. While recent elicitation studies have targeted written-mode MLP, factors underlying the MLP construct across written-/spoken-modes are unknown. We address these issues via two research questions (RQs):

1. What are the latent relationships between written/spoken/receptive/productive MLP measures for advanced L1 Mandarin L2 English learners?
2. Which linguistic/conceptual/metalinguistic themes characterize the elicited MLP?

For comparability with O'Reilly and Marsden (2021), we investigate L2 MLP in higher-proficiency L1 Mandarin L2 English learners rather than across L1s/L2s/L3s, and so on, which present as useful future research directions.

Method

All data collection materials/data/analysis scripts are available on the study's Open Science Framework (OSF) page (https://osf.io/tv9eg/) and via the IRIS database (www.iris-database.org).

Participants

Participants were 69 L1 Mandarin L2 English speakers (61 females, 8 males) aged 21–35 ($M = 24.83$, $SD = 2.74$), of 'advanced' proficiency having passed the criterion-referenced Test for English Majors (TEM) 8 (Jin and Fan 2011) in recent years.[1] Most ($n = 41$) had completed

their undergraduate degree and were using English in high-school teaching (English, Art), English-Mandarin translation, international sales, business, administration, data analysis, or HR. Others (n = 23) were undertaking postgraduate study in English Education, Translation and Interpretation, Legal Translation, Business English, and Management. The remaining five graduated from their undergraduate English major degree around the time of data collection (June 2019).

Data collection instruments

The MLP written-mode and spoken-mode receptive and productive tests, their origin, item/response format, scoring, and foci are summarized in Table 1.

MLP written-mode receptive and productive tests

These measures were constructed using Metaphoric Competence test items from O'Reilly and Marsden (2021) which (in that study) were retained after various data cleaning steps (e.g. item analysis) and loaded on the MLP factor. These items were designed to tap receptive and productive idiom extension/re-literalization (Littlemore and Low 2006) and receptive and productive metaphor continuation (Low 1988). Respective examples (current study Q2, Q7, Q4, and Q10) are shown in Figure 1.[2]

Responses to the MLP written-mode receptive test (hereafter MLP written-mode-R) were scored 1 (successful) or 0 (unsuccessful), and to the productive test (hereafter MLP written-mode-P) scored 2 (successful), 1 (partially successful) or 0 (unsuccessful).[3] For all written-mode tests, the groups of three items were preceded by instructions and an example.

All the current study's items targeted comprehension/production of either a figurative or re-literalized/blended (hereafter 'literal') meaning. To understand the nature of successful/less successful responses, we further coded comprehensions/productions as predominantly concerning a figurative or literal meaning and whether they displayed higher or lower success. For example, Q2 in Figure 1 elicited comprehension of a literal meaning best-answer 'come back with a walking stick!', scoring 1, coded 'literal successful' as an appropriate re-literalization of the phrase (as per the task instructions). Other options were designed as distractors, scoring 0 because of subtle issues: 'see where you can break your leg!' (re-literalized but less successful, the reason for focusing on location is unclear); 'do the very best you can!' (successful as a figurative meaning but this is not what the task elicited); and 'do something that gets you injured!' (literal but less successful, injury is already assumed, an extension is needed). Productive responses were coded in the same way. Given our scoring focus was 'communicative success', spelling/grammar issues were not penalized in scoring provided the meaning could be reasonably understood, but were analysed and discussed thematically as evidence of MLP development (RQ2).

MLP-spoken-mode receptive and productive tests

These constructs were operationalized via a specially developed test (hereafter MLP-spoken-mode-R) eliciting comprehension of a 54-s video clip from Season 1, Episode 1 (1994) of the US sitcom *Friends*, analysed by Dore (2015) and a related production test (hereafter MLP-spoken-mode-P). Piloting revealed comprehension difficulties and participants' wishes to include English subtitles, a decision which we reflect on in Discussion and Limitations. For convenience, we refer to this as a 'spoken-mode' test but make clear that it involved listening/viewing the clip, reading subtitles, and demonstrating comprehension to the interviewer via speaking. MLP-spoken-mode-P involved listening to the researcher's questions and providing oral responses.

Table 1 shows MLP-spoken-mode-R questions, their target meaning, and scoring, which was developed using a rubric developed from Dore's (2015) analysis, pilot responses, and a sample of main data. Q14 served only to support Q13 to ensure participants had understood the main meaning before proceeding and so only descriptive statistics are reported for this item; it was not used in composite scores or the Exploratory Factor Analyses.

Table 1: MLP written-mode tests (adapted from O'Reilly and Marsden 2021) and spoken-mode tests (specially developed)

Test[a]	Item format	Target response meaning[b]	Scoring (further coding)[c]	Question (Q), test of ability to:
MLP-written-mode-R				
	Multiple-choice (four options)	Literal	1(L+), 0(L–/F–)	Q1. Recognize extension of the literal senses of idiom: 'It's been raining cats and dogs for so long that…!'
	Multiple-choice (four options)	Literal	1(L+), 0(L–/F–)	Q2. Recognize extension of the literal senses of idiom: 'Just go out and break a leg. In fact, go out and…!'
	Multiple-choice (four options)	Literal	1(L+), 0(L–/F–)	Q3. Recognize extension of the literal senses of idiom: 'We were so stuck between a rock and a hard place that…!'
	Multiple-choice (four options)	Figurative	1(F+), 0(F–)	Q4. Recognize continuation of metaphors for pregnancy in discourse 'So you're telling me that…?'
	Multiple-choice (four options)	Figurative	1(F+), 0(F–)	Q5. Recognize continuation of metaphors for pregnancy in discourse: 'What about gender? Will you…?'
	Multiple-choice (four options)	Figurative	1(F+), 0(F–/L+)	Q6. Recognise continuation of metaphors for pregnancy in discourse 'I'm so glad to hear that once again you'll be…!'
MLP-written-mode-P				
	Gap-fill	Literal	2(L+), 1(L–), 0(L–/F+/F–/N)	Q7. Produce extension of the literal senses of idiom: 'He beat around the bush for so long that…!'
	Gap-fill	Literal	2(L+), 1(L–), 0(L–/F+/F–/N)	Q8. Produce extension of the literal senses of idiom: 'In fact, it didn't just take the cake, it…!'
	Gap-fill	Literal	2(L+), 1(L–), 0(L–/F+/F-/N)	Q9. Produce extension of the literal senses of idiom: 'Please cross that bridge when you come to it. Although, since the decision seems likely, my advice is to…!'
	Gap-fill	Figurative	2(F+), 1(L+/F–), 0(F–/L–/N)	Q10. Produce continuation of metaphors for an award-winning colleague in discourse 'I heard Mr Magic is due to be…?'
	Gap-fill	Figurative	2(F+), 1(L+/F–), 0(F–/L–/N)	Q11. Produce continuation of metaphors for an award-winning colleague in discourse: 'Will the magic circle commend him for…?'
	Gap-fill	Figurative	2(F+), 1(L+/F–), 0(F–/L–/N)	Q12. Produce continuation of metaphors for an award-winning colleague in discourse 'I agree, I'm completely…!'
MLP-spoken-mode-R (+subtitles)				
	Verbal explanation	Figurative	2(F+), 1(F+), 0(F–/L–)	Q13. Interpret main meaning: 'What do you think is the main meaning of this conversation?'

Table 1. Continued

Test[a]	Item format	Target response meaning[b]	Scoring (further coding)[c]	Question (Q), test of ability to:
	Verbal explanation	Figurative	1(F+), 0(F–/L–/N)	Q14. Interpret ice-cream metaphor: 'Joey (one of the characters) talks about different flavours of ice cream. What does this mean?'[d]
	Verbal explanation	Figurative	2(F+), 1(F–), 0(L+/L–/F–/N)	Q15. Interpret spoon metaphor: 'Joey also says "grab a spoon". What does this mean?'
	Verbal explanation	Literal	2(L+), 1(F-/L-), 0(F-/N)	Q16. Interpret hungry/horny joke: 'At the end of the conversation, Ross says "I don't know if I'm hungry or horny". What is the meaning here?'
	Verbal explanation	Literal	2(L+), 1(F–/L–), 0(F–/L-/N)	Q17. Interpret freezer joke: 'Chandler (the other friend) says then "stay out of my freezer". What is the meaning here?'
MLP-spoken-mode-P				
	Verbal production	Figurative	2(F+), 1(F–), 0(F–/L+)	Q18 Produce suitable ice-cream-self metaphor '(a) Romantically, or perhaps in your friendships, do you think that you are a kind of ice cream? Why? (b) If you are ice cream, what could you say about the different features (for example, flavour, colour, taste, price, texture)? Why?'[e]
	Verbal production	Figurative	2(F+), 1(F–), 0(L+)	Q19 Produce suitable other-self metaphor: '(a). Except ice cream, would you like to use anything else (animal, plant, accessories or anything else) to describe yourself (for example, I am a cat or I am a flower)? Why do you think in this way? (b) If you are a [participant's answer], what are the features of it? (Material? Colour? Etc.)'[e]
	Verbal production	Figurative	2(F+), 1(F–), 0(F–/L+)	Q20. Produce suitable food-people metaphor: 'We may talk about different people as if they were different foods. How about your friends or families, what kinds of food are they? Why?'
	Verbal production	Figurative	2(F+), 1(F–), 0(F–/L+/N)	Q21. Produce suitable food-life/other metaphor: 'We can also talk about life as food. What kind of food or anything else is your life? Why?'

[a]-R = receptive knowledge test, -P = productive knowledge test.
[b]literal = re-literalized/blended (see 'Data collection instruments' section).
[c]This column shows further coding of comprehensions and productions within the various scores, F+ = figurative successful, F– = figurative less successful, L+ = literal successful, L– = literal less successful, N = no response.
[d]Item to ensure all participants received explanation of metaphor, not used in any analyses involving composite variables or the Exploratory Factor Analysis.
[e]Questions (a) and (b) asked sequentially but a single score awarded.

Finally, MLP-spoken-mode-P elicited playful productions of the same PEOPLE ARE FOOD metaphor and other related metaphors. The questions, their target meaning, and their scoring can be seen in Table 1. In a few cases where participants declined to produce a metaphor due to the pejorative undertones (comparing women/oneself to ice cream, as in the *Friends* clip), we took this as evidence of their awareness of a socially sensitive metaphor (Low 1988) and scored the non-production as 2, a communicatively successful resistance of the metaphor. Interestingly, such participants were generally happy to then compare friends/family members and their lives to foods and other entities.

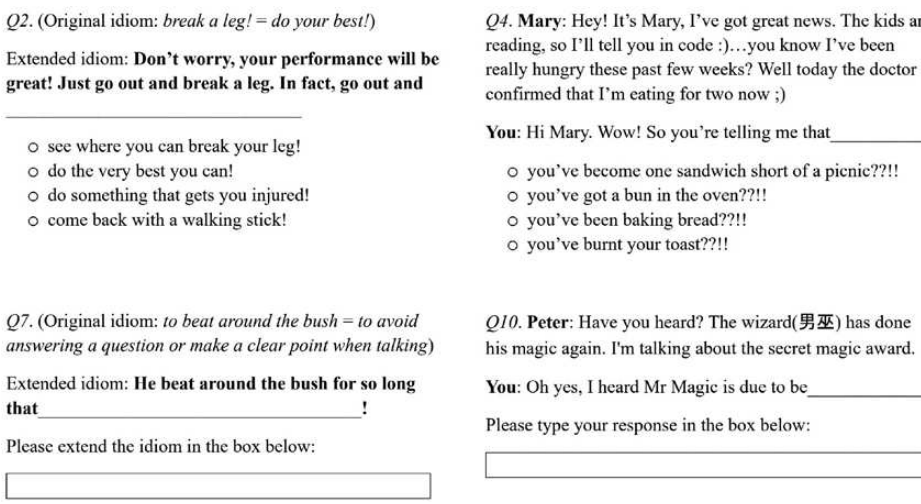

Figure 1: Examples of MLP written-mode test items. Receptive and productive items are shown on the top and bottom row, respectively, idiom extension and metaphor continuation on the left and right side, respectively.

Procedure

Before main data collection, we piloted draft tests with four TEM-8 qualified participants (three female, one male) of similar age who were also asked to think aloud. As a result, to reduce the test-taker burden, MLP written-mode-R and -P test items were reduced by half, from the full 12 used in O'Reilly and Marsden (2021) to the most highly scoring items: three receptive and three productive idiom extension/re-literalization items, and three receptive and three productive metaphor continuation items. Where unfamiliar vocabulary might have impeded MLP, several multiple-choice options were paraphrased (e.g. Q2. 'crutches' changed to 'walking stick') and two Mandarin translations added ('wizard [男巫]', 'spells' [咒语]). After the authors had discussed and agreed scores, an additional rater blind-scored 10 per cent (105/1035) of all non-multiple-choice items, with inter-score reliability showing 'substantial' agreement (weighted kappa = 0.85) (Landis and Koch 1977). The remaining disagreements were resolved through discussion.

Main data collection took place in June 2019 in individual, synchronous online sessions. To control for possible order effects, three groups of 23 participants, balanced for TEM-8 score and gender, completed MLP written-mode -R and -P, MLP-spoken-mode-R and -P and a third task not relevant to the current study's RQs (Yan 2019) in different orders.

MLP written-mode-R and -P were completed online via Qualtrics link. The *Friends* clip was emailed, participants viewed it, and answered an ice-breaker question about whether they had seen/liked the series, followed by main questions. Participants watched the clip a second time before proceeding to MLP-spoken-mode-P. Here, if the same metaphor was repeated, this was credited only once. If participants could not think of an answer the researcher prompted with ideas; in only 2 out of 276 responses (0.7 per cent), a participant was unable to think of any metaphor. Participants did not generally receive feedback, but in a few cases (13/276 responses, 4.7 per cent), the researcher confirmed the answer to motivate and allay apparent concerns, minor interventions unlikely to have impacted on substantive findings. Participants were thanked and emailed feedback to support their future English learning.

Data analysis

Data were analysed using R programming language (R Core Team 2022) with various packages/scripts and Microsoft Excel.

First, MLP written-mode-R and -P responses were downloaded into MS Excel and audio-recorded MLP-spoken-mode-R and -P responses were transcribed verbatim with square parentheses indicating the researcher's dialogue. To address RQ1, descriptive statistics of item and composite scores are first reported to summarize areas of ease/difficulty in the measures variables and contextualize subsequent analyses, followed by an Exploratory Factor Analysis to uncover latent MLP relationships between measured variables (Loewen and Gonulal 2015). In other words, we first sought to understand where participants were succeeding/struggling on the various tasks before discovering the underlying competences that were, to varying degrees, responsible for this. Exploratory Factor Analysis was chosen in favour of a Rasch-family approach (Knoch and McNamara 2015), since simulation studies have shown factor analytic approaches excel at identifying multiple factors where the objectives align with ours, namely, to explore the underlying concepts that items are measuring (Chen *et al.* 2014a), where there is no prior knowledge of a dominant factor (Chen *et al.* 2014b), and in cases where only small correlations between factors exist (Smith 1996, cf. Christensen *et al.* 2012).[4]

Before the factor analysis, we imputed 2/1449 (0.1 per cent) missing scores where Q20 was skipped, using the MICE package in R (see Supplementary material). Following methodological guidance (Field 2013), four variables with Kaiser–Meyer–Olkin values below 0.5 (Qs4–6, 18) were removed to maximize data factorability, resulting in a 16-variable correlation matrix, adequate for factor analysis by various preliminary and post hoc indicators (see Supplementary material).

For the main Exploratory Factor Analysis, we retained a three-factor solution as the most parsimonious/informative model, balancing stricter and more liberal factor retention criteria (Loewen and Gonulal 2015; Plonsky and Gonulal 2015; O'Reilly and Marsden 2021). Principal axis factoring, entailing no distributional assumptions, mitigated the impact of uni- and multivariate nonnormality (Fabrigar *et al.* 1999), while Direct Oblimin (oblique) rotation accounted for possible interrelatedness between factors. As in O'Reilly and Marsden (2021) factor loadings above 0.32 were interpreted (Tabachnik and Fidell 2013), and when interpreting factors we carefully considered the ability each loading item was measuring and looked closely at pattern matrix loading strengths and the likely replicability/stability of factors as indicated by the lower 95 per cent confidence interval of 5,000 bootstrap resamples (Zientek and Thompson 2007) (see Supplementary Table S1).

Item-within-factor internal consistency was estimated using ordinal omega (ω) due to ordinal data, item-test construct relationship variation (congeneric scales), and the equal weighting of items in overall test score calculation (McNeish 2018; see also O'Reilly and Marsden 2021). Estimates were interpreted in relation to those obtained from the Metaphoric Competence Test Battery (O'Reilly and Marsden 2021) and a methodological synthesis of instrument reliability in L2 research (Plonsky and Derrick 2016). Given these checks, and since all instruments were piloted and written-mode items sourced from the optimized MC Test Battery (O'Reilly and Marsden 2021), we did not employ an item analysis approach in the current study. Thus, indications of item ease/difficulty are provided by the various descriptive statistics.

To address RQ2 all 1,449 comprehensions/productions were thematically analysed (Cohen *et al.* 2017) via coding for 'elicited_meaning [of the item]' (figurative, literal), 'response_code' (figurative successful, figurative less successful, literal successful, or literal less successful), and 'theme' (a description of linguistic/conceptual/metalinguistic properties), with accompanying 'notes' (explanation of the score or theme). In the 'Results and Discussion' section key themes and examples are reported with supporting frequencies/proportions (for the full set of results, see Supplementary Table S2).[5]

Results and discussion

For readability and concision, we present results for RQ1 and RQ2 with integrated discussions in this section. Shifts from results to discussion are indicated by the presence of interpretations, comparisons with previous research, and citations.

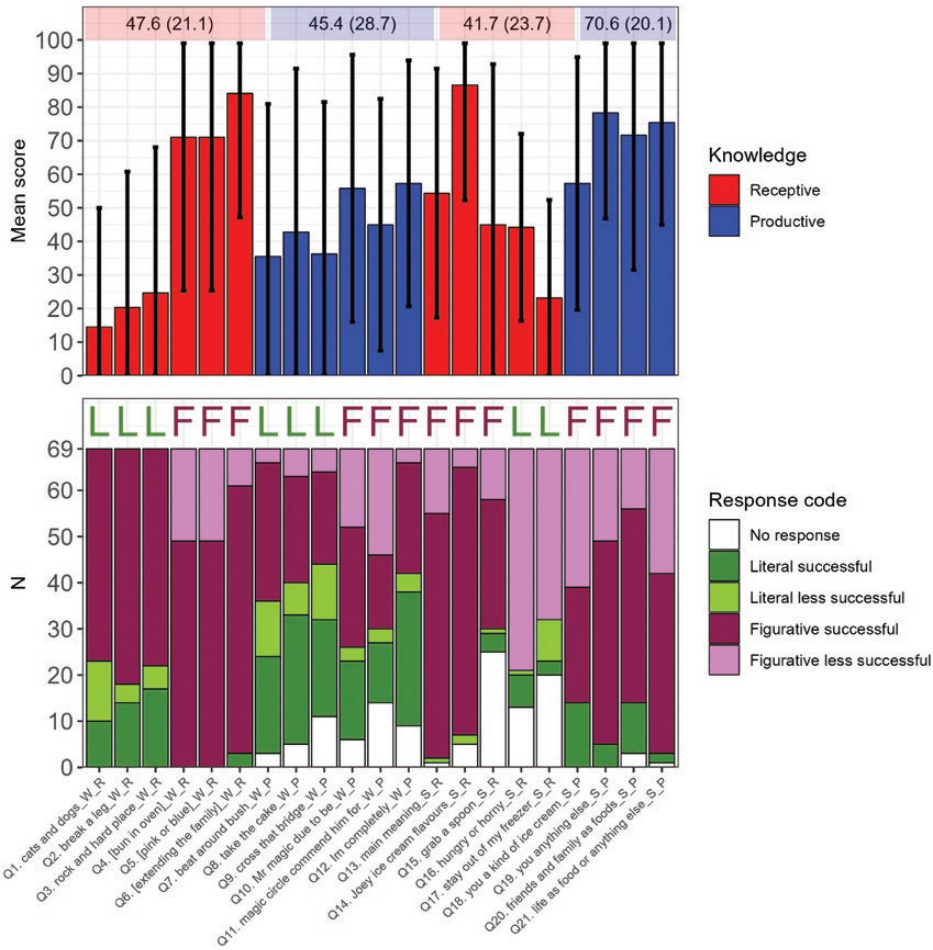

Figure 2: Both plots: N = 69, x-axis = item (see also Table 1). Top plot: At the top, left to right, overall MLP written-mode-R (receptive) and -P (productive), and MLP-spoken-mode-R (receptive) and -P (productive) mean scores (standard deviations) excluding Q14, y-axis = mean item difficulties as percentage scores with standard deviations (black bars); Bottom plot: y-axis = MLP counts per response type for items eliciting target figurative (F) or literal (i.e. re-literalized/blended) (L) meaning comprehension/production, all responses assigned one of four codes (see legend and 'Data collection instruments' section).

RQ1: Latent relationships between learners' written/spoken/receptive/productive MLP

Figure 2 (top plot) shows that participants' (Ps') MLP-spoken-mode-P scores were highest ($M = 70.6$, $SD = 20.1$), that is, they overall excelled at verbally generating playful comparisons involving conceptual mappings and elaborations anchored in one main metaphor, for example, '[Q20] my boyfriend [is like] a durian because … (P11)'. This is unsurprisingly given that MLP-spoken-mode-P items (Qs18–21) elicited relatively straightforward X-is-Y-because-Z formations rather than complex re-literalizations (cf. Qs1–3, 7–9, 16–17), the richness of food imagery, and the personal nature of MLP-spoken-mode-P items compared with others involving (unknown) characters. Overall scores for MLP written-mode-R ($M = 47.6$, $SD = 21.1$), MLP written-mode-P items ($M = 45.4$, $SD = 28.7$), and MLP-spoken-mode-R items ($M = 41.7$, $SD = 23.7$) were notably lower.

Concerning individual items, excluding Q14, scores were highest for Q6-[extending the family]-W-R ($M = 84.1$, $SD = 36.9$) and high for receptive metaphor continuation items (Qs4–6) generally, perhaps due to well-known imagery linked to pregnancy ('bun in the oven', 'pink or blue'). For these items, the current study's advanced learners were better able to select the most successful option and reject distractors compared with mixed-proficiency learners in previous research (O'Reilly 2017). Taken together, these points are useful in pinpointing areas of relative ease for different levels of learner, before any pedagogical intervention has taken place, and thus, for identifying the most accessible types of MLP for English language teaching (Macarthur 2010).

Analysis of item ease/difficulty by target meaning revealed several nuances. In the written-mode, where items elicited a target literal meaning (Figure 2, bottom plot), producing idiom re-literalizations (Qs7–9) was easiest (21–28/69 'literal successful' productions) and, interestingly, comprehending them (Qs1–3) most difficult (10–17/69 'literal successful' selections), a finding in line with previous research (O'Reilly and Marsden 2021). In the spoken-mode, comprehension of the re-literalized meanings (Dore 2015) in Q17-stay out of my freezer-S-R and Q16-hungry or horny-S-R was particularly challenging (3/69 and 7/69 'literal successful' responses, respectively) whereas target figurative meanings, producing metaphor continuations (Qs10–12) was most difficult (16–26/69 'figurative successful').

Turning to the latent dimensions, the Exploratory Factor Analysis pattern matrix (Figure 3) shows a three-factor solution for the 16 variables, explaining 36 per cent total variance, 8–20 per cent variance for individual factors.

The factors indicate three underlying MLP skill areas. Factor 1 (ordinal $\omega = 0.94$) was named 'Idiom Re-literalization in Written Tasks' since its five loadings concerned receptive/productive idiom re-literalization (O'Reilly and Marsden 2021). Factor 2 (ordinal $\omega = 0.71$) was named 'MLP (Comprehension) in Spoken Dialogue' since its five loadings predominantly concerned receptive spoken-mode items. Factor 3 (Spearman-Brown coefficient = 0.65, see Eisinga *et al.* 2013) was named 'Metaphor Continuation in Written Interaction' since its two loadings concerned productive knowledge of metaphor continued throughout a written discourse (O'Reilly and Marsden 2021). The lower bound 95 per cent confidence intervals from bootstrap loadings were positive (>0) for most (10/12) loadings indicating a stable and likely replicable solution (Zientek and Thompson 2007), particularly for factor 1, for which all lower bounds were also above .32.[6]

Two main substantive findings emerge from the Exploratory Factor Analysis. First, within our set of written/spoken/receptive/productive MLP measures, written-mode idiom extension/re-literalization and metaphoric continuation load on different (but related) factors, whereas in previous research, they were shown to both relate to an MLP factor *within* wider Metaphoric Competence (O'Reilly and Marsden 2021). This suggests distinct MLP abilities falling within the same general Metaphoric Competence area of MLP. Since vocabulary depth is a strong predictor of overall MLP (O'Reilly and Marsden 2023), future research here might explore the extent to which its explanatory power extends to the three MLP factors identified.

Second, the current study's three MLP factors generally affirm a distinction between written-/spoken-modes and receptive/productive knowledge. However, modes were more disparate since the two written-mode factors (Idiom Re-literalization in Written Tasks and Metaphor Continuation in Written Interaction) were positively correlated, but neither related to MLP (Comprehension) in Spoken Dialogue. Only Metaphor Continuation in Written Interaction was completely homogenous with respect to receptive/productive knowledge, suggesting high relatedness of these two knowledge types. By implication, executing one's associative thinking abilities (Littlemore 2001; Littlemore and Low 2006; Littlemore 2008; O'Reilly and Marsden 2023) during more fluid, working memory-intensive spoken-mode MLP does not seem to relate well to doing so with more static, less working memory-burdening written MLP, and vice versa (O'Reilly and Marsden 2021). However, these points require corroboration, since differences between the current study's written-/spoken-mode item foci and the use of subtitles in MLP-spoken-mode-R test complicate a neat interpretation.

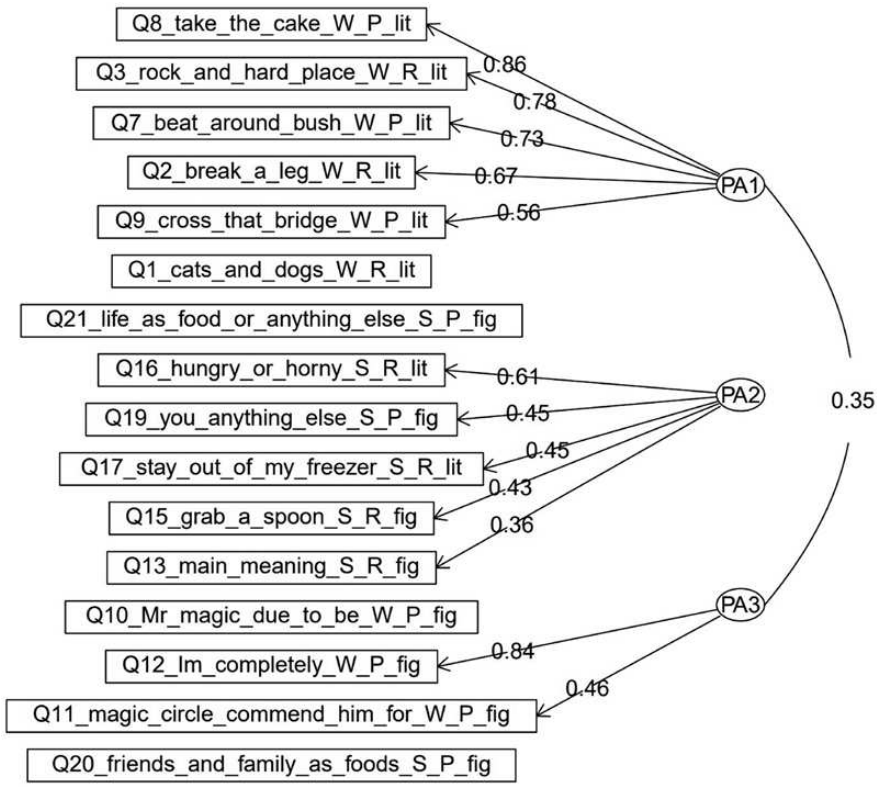

Figure 3: Diagram of MLP factors with loadings <.32, Direct Oblimin rotated (36 per cent total variance explained), PA1 = factor 1 (20 per cent), PA2 = factor 2 (8 per cent), PA3 = factor 3 (8 per cent), W/S = writing/speaking, R/P = receptive/productive, fig/lit = predominantly figurative/literal meaning elicited.

Methodologically, MLP item-within-factor internal consistency was reasonable (*Mdn* = 0.71, IQR = 0.15), above the 25th percentile of 23 estimates from L2 Metaphoric Competence research (0.69), equivalent to the median of 14 Metaphoric Competence tests (0.71) and 25th percentile of four subsequent Metaphoric Competence factors reported in O'Reilly and Marsden (2021), but just below the 25th percentile from a methodological synthesis of 1323 estimates in L2 research (0.74) reported in Plonsky and Derrick (2016). This finding further supports the notion that Metaphoric Competence measures show comparatively high measurement reliability variability compared to the wider L2 field.

RQ2: Linguistic, conceptual, and metalinguistic themes in learners' MLP

In this section, we present a selection of the most salient linguistic, conceptual, and metalinguistic themes in the response data with supporting examples. For a bird's eye view of the thematic analysis, including frequencies/percentages of themes grouped by score, response code, and item, readers are referred to Supplementary Table S2.[7]

Factor 1: Idiom Re-Literalization in Written Tasks

This factor was most strongly marked by the Q8-take the cake_W_P loading (0.86), for which 41 per cent (28/69) of productions scored 2/2, 4 per cent (3/69) scored 1/2, and 55 per cent (38/69) scored 0/2. Most successful productions were characterized by their invoking of a humorous,

hyperbolic re-literalized situation effect, for example: [his comment, didn't just take the cake it] 'ruined the whole party (P12)', 'won a dessert shop (P63)', 'made a mess of the table (P69)'. P12 and P69's extensions both use disaster, whereas P63 adopts a more unusual play with the idiom's potentially positive meaning (i.e. the best/worst of its kind). The going-one-step-further theme employed both food: 'took the cake away (P3)', 'takes all the desserts (P13)', 'also took the champagne and flowers (P34)'; and objects: 'took the candles (P18)', 'take the cake pan, too (P21)', 'even took the box as well (P66)'. At the linguistic level, while 20/28 of the successful responses were grammatical given the stimuli, 8/28 were ungrammatical, containing a present tense verb form with third person -s: 'takes (P13, P43)', 'robs (P19)'; without third person-s: 'take (P21, P23, P50)'; an incorrect regular -ed: 'lighted (P45)'; and a misspelling of *eat/ate*: 'eta (P68)', difficulties likely explained by the typological distance between English and Chinese (Murakami and Alexopoulou 2016).

For Q7-beat around bush_W_P, another strong loading item (0.73), 30 per cent (21/69) of productions scored 2/2, 10 per cent (7/69) 1/2, and 59 per cent (41/68) 0/2. Most successful extensions played with an environmental effect on the bush and surroundings, on occasion with opposite but equally successful meanings of fright and curiosity/hunting: [he beat around the bush for so long that] 'all animals in the bush run away (P17)' versus 'he just attracted wild animals (P42)'. Other successful responses were character-effect focused: 'he almost broke his stick (P12)', 'he was dizzy (P23)', and 'he can not walk anymore (P43)'. Linguistically, differences in correct/incorrect tense are observable in P12's and P23's versus P43's responses.

With Idiom Re-literalization in Written Tasks productive items (Q7–9), partially successful responses (scoring 1/2) were consistently infrequent compared with successful or unsuccessful responses, suggesting something of a polarization of productive knowledge on this factor. Partially successful productions were somewhat logical with unclear aspects, as in: [Q8 … his comment] 'cut into the point (P25)', where the (literal or figurative?) point confuses what might have been intended as cutting into the cake; 'is the cake (P32)', where the significance of the comment becoming the cake is unclear; and 'lit fiweworks (P65)', where it is unclear whether fireworks or candles were intended.

Unsuccessful productive item responses were most commonly extensions that successfully extended the figurative (not literal) sense: [Q7] 'he don't want to tell me the answer clearly (P27)' and 'no one gets his idea (P10)'. Following that were responses that less successfully extended the figurative sense: [Q7] 'everyone understands him (P14)' and 'we all understand his meaning (P24)'; and indecipherable extensions of literal meanings: [Q7] 'he hid himself for a long time (P25)', the reason why being unclear, and 'he forgot his prey (P38)', where the character's reason for hunting is unclear. Here, P22 repeated part of the item stimulus (the original idiom): [Q7] 'He is unwillig to answer, and just beat arond the bush (P22)', a strategy evident in their other idiom extensions, for example, Q9-cross that bridge_W_P (factor loading 0.56), 'I worry, you can handle this when you cross the bridge when you come to it (P22)'.

In sum, while many participants had a good conceptual knowledge of the boundaries of creative acceptability (Low 1988), with productive MLP generally, the variability of phraseological proficiency/linguistic accuracy is notable (Philip 2010). These findings extend previous research on learner malformations with collocations (Miller 2011) and idioms (Philip 2010), providing evidence of grammatical inaccuracy in otherwise successful creative idiom extensions/re-literalizations and the nature of inaccuracies, which for these learners mostly concerned stimuli-response tense agreement, likely attributable, in part, to cross-linguistic influence (Murakami and Alexopoulou 2016). These points suggest that pedagogical activities aimed at developing learners' MLP should pay sufficient attention to both conceptual creativity and linguistic accuracy (Boers 2000; Philip 2010) to maximize (figurative) interlanguage destabilization, fossilization prevention, and more holistic competence development (Tarone 2005; Macarthur 2010; Bell 2012a, 2012b).

Idiom Re-literalization in Written Tasks also comprised two receptive items, Q3-rock and hard place_W_R (factor loading 0.78) and Q2-break a leg_W_R (factor loading 0.67). Two trends emerge when these (advanced) participants' responses are compared with mixed-proficiency learners from previous research who completed these same items (O'Reilly 2017; O'Reilly and Marsden

2021). First, and somewhat confusingly, our participants were more easily lured from the successful re-literalization extension, best answer by the figurative successful distractor: for Q3-rock and hard place_W_R, 68 per cent (our study) versus 38 per cent (in previous research) selected 'we were getting very worried!', for Q2-break a leg_W_R, 74 per cent versus 43 per cent selected 'do the very best you can!'. This may be because the current study's learners had stronger form-meaning idiom knowledge links, drawing them to the idiom's salient/figurative meaning. Alternatively, our participants may have been less attentive to the instruction example re-literalization compared with learners completing this test within the Metaphoric Competence Test Battery (O'Reilly 2017; O'Reilly and Marsden 2021), treating the task as spot-the-idiom-meaning (O'Reilly 2017). However, given the numerous attempted idiom re-literalizations in production (Qs7–9), it may be that compared with multiple-choice, the limited-production gap-fill enables and encourages more risk-taking with this kind of MLP.

Second, our participants were comparatively better at rejecting the two less successful re-literalization distractors than in previous research: for Q3-rock and hard place_W_R, 6 per cent versus 7 per cent selected 'we were falling into the ground!' and 1 per cent versus 9 per cent 'our feet were going soft!' (total 7 per cent vs. 16 per cent), for Q2-break a leg_W_R, 4 per cent versus 7 per cent selected 'see where you can break your leg!' and 1 per cent versus 11 per cent 'do something that gets you injured!' (total 6 per cent [rounded] vs 18 per cent). Since idiom re-literalizations reside on the novel/non-stock phrase end of the spectrum (O'Reilly and Marsden 2021), guesses at appropriately creative MLP (Low 1988) probably reflect successful associative and figurative thinking (Littlemore 2001; Littlemore and Low 2006; Littlemore 2008; O'Reilly and Marsden 2023). However, further research is needed to ascertain such cognitive processes and the wider question of how learners develop knowledge of creative acceptability boundaries.

Despite the challenges of receptive re-literalization, the multiple-choice items have potential pedagogical uses and benefits. For example, because form recognition offers learners the opportunity to analyse different options, all else being equal this format should, theoretically, be easier than explain-the-meaning/meaning recall (Laufer and Goldstein 2004) and promote deep processing (Craik and Lockhart 1972), a consideration that practitioners might bear in mind when navigating MLP with underconfident learners. Also, compared with open gap-fill/limited production, multiple-choice gives learners and teachers MLP content/input to work with before progressing to less restricted tasks, which could pose problems if administered prematurely in the learning sequence (Low 1988). Furthermore, distractors such as those discussed above help pinpoint where learners struggle in rejecting less successful language and enable rich and stimulating classroom discussion about what makes for 'good/bad' MLP from general and local L2 English, ELF, and multilingual perspectives (Pitzl 2016). Although learners sometimes inadvertently (mis)acquire less successful metaphorical language input (e.g., Boers *et al.* 2014), negative evidence of non-target-like MLP is key to helping learners notice and develop knowledge of what speakers of the target variety tend *not* to say (Low 1988; Trahey and White 1993).

Factor 2: MLP (Comprehension) in Spoken Dialogue

This factor was most strongly marked by the Q16-hungry or horny_S_R loading (0.61). Only 10 per cent (7/69) of participants successfully comprehended Ross's joking figurative/literal ice cream-dating imagery conflation, for example: 'it might be I'm hungry because talking about food using food as a metaphor but the horny part because I think … there's kind of like a picture in Ross' mind instead of ice cream he's thinking of different things he can do with them [women] (P68)'.

Most partially successful interpretations scoring 1/2 (67 per cent, 46/69) focused exclusively on the figurative meaning of the metaphor (physical/romantic attraction to women), in particular on Ross' anxiety/uncertainty about what to do, as in: 'maybe he just don't know whether he needs women right now (P4)'. Conversely, other responses identified Ross' plan going forward, for example: 'hmm maybe he's eager to date or to marry a new wife or girlfriend (P38)'. One response, describing Ross as looking back rather than forward, contained a possible strategic coinage for 'ex-wife', 'he cannot forget the pre-wife (P1)'. Just one participant focused exclusively on the literal meaning

of ice cream as food, '… since his friend been talking so many flavours of ice cream he think about flavour maybe I should think some ice cream now (P64)'. Unsuccessful responses, scoring 0/2 (22 per cent, 15/69), showed clearly unsuccessful comprehension, for example: 'I'm hungry or horny I think it looks like a good guy I think (P23)'.

Just 4 per cent (3/69) of participants successfully comprehended another key loading item, Q17-stay out of my freezer_S_R (0.45), offering tentative explanations: '… Ross will … how to say, can treat the freezer like a lady (P28)', '… maybe he did something sexual when he opens the fridge, so it's kind of humour's way to encourage Ross or make the situation less embarrassed (P34)' in which a pragmatic inference is also offered, and '… maybe … he says stay away with my freezer it's like stay away from me, I'm not your ice cream (P63)'. After discussion between scorers, P63's interpretation was deemed successful because they recognized the implied threat to Chandler and his interests more generally (cf. Dore 2015).

For this item's unsuccessful responses, comprising 58 per cent (40/69), the most common theme was participants' explicit recognition of their non-understanding. Where an interpretation was attempted, there was some evidence of figurative thinking around individual lexical components (Littlemore and Low 2006), for example, 'a freezer is used for storing ice cream so out my freezer means just like I can eat them, right? Maybe means he doesn't like those women, that ice cream, their understandings of that, the one is that's not my favourite ice creams I put them away I don't like them, the second is they are out of the freezer they will come to my mouth (P2)'. Here, P2 first links 'freezer' and 'out' to posit ice cream ready for consumption, offering the alternative misinterpretation that Chandler dislikes a particular kind woman or ice cream, before returning to the incorrect notion of ice cream ready for consumption.

The central pedagogical implication arising from these observations is for educators to help L2 learners become aware of the possibility of concurrent, interacting, and blended meanings when navigating sitcom humour (à la Lucas 2005; Tocalli-Beller and Swain 2007) to enable subsequent form-meaning comprehension and uptake (McDaniel *et al.* 1995; Schmidt and Williams 2001; Strick *et al.* 2010). If learners can notice and comprehend MLP in real-world language, this may promote increased risk-taking in their production. Although some learning gains have been observed following awareness-raising interventions directing attention to the conceptual underpinnings and etymological elaboration of phrasal verbs, unamended idioms, and collocations (Boers 2000; Boers *et al.* 2007), the extent to which such approaches might help MLP develop is unknown. While researchers are generally suspicious of 'magic bullet' teaching/learning solutions to Metaphoric Competence development (Macarthur 2010: 158; see also Nacey 2013), the above strands of literature and current data provide a basis for identifying, testing, and refining ways of improving MLP.

As a possible pedagogical application of MLP (Comprehension) in Spoken Dialogue items, a teacher might, for example, use the *Friends* clip as the key language input and basis for a lesson on the topic of 'break-ups'. For higher-proficiency learners, the lesson could progress from a lead-in/brainstorming of romantic break-up vocabulary to listening/viewing the clip for gist (slowed down or with subtitles, if appropriate), and then for specific information to unpack the various metaphor components (e.g. providing the current study's successful and unsuccessful responses on what ice cream flavours stand for as multiple-choice options), to freer practice. In the latter stage, small groups might be assigned/choose an idiom (e.g. plenty more fish in the sea, every cloud has a silver lining) and fictional characters, and pre-plan/improvise a dialogue in which they must playfully extend the idiom to console a friend experiencing a break-up (e.g. '… the sea is a big place!'), who does not want to move on ('… there's only one fish for me!').

Another pedagogically relevant theme in this factor's response data was the various metalanguage. One way this could be seen is in how participants refer to metaphor itself, either explicitly 'metaphor (P68)', or via other means such as 'the deeper meaning (P19)' and a 'rhetoric device (P20)'. Another manifestation of metalanguage, seen in Q13-main meaning_S_R (0.36) responses, was in how participants explained the workings of the metaphor in question. Here, the source

(ICE-CREAM) and target (DATING) link was explained using a range of terminology for indirect metaphor/simile/comparison: 'compare the woman to different flavours (P43)', 'women is just like ice cream (P16)', 'taking … the flavour of ice cream as an example to tell this man that there are so many women in the world (P1)'. Other main meaning interpretations specified the term 'metaphor': 'I think Joey is trying to tell something about the men and women in the relationship by using the metaphor of the ice cream flavours (P15)'; used X is/means Y formulations: 'different flavours means different kinds of women (P50)'; and even the characters' metaphor itself: 'two of his friends may be talking to him and to encourage him to go out from his divorce because there's many flavours, many other women for him not just her ex (P56)'. Metalanguage, however, was not always successfully recalled: '[Q16-hungry or horny_S_R] it's kind of similar English similation, assimilation I forgot the terminology (P37)'.

These points have several implications for explicit MLP instruction and presentation, where educators/materials designers may have pre-conceived ideas about the most accessible and helpful terminology to prescribe or employ. The data show that although some learners explicitly talk about metaphor as 'metaphor', others use different, more personalized metalanguage to conceptualize this phenomenon. While previous (M)LP research has shown how learners negotiate new vocabulary meanings with interlocutors during language-related episodes (Bell 2005; Pitzl 2016), the current study provides insights into how learners react to and conceptualize MLP input *without* interlocutor assistance. For educators, the key consideration is finding an appropriate balance between helping learners talk about language in ways that work for them and teaching more widely recognized metalanguage and terminology. Here, further research could usefully draw on methodologies used to elicit and investigate metaphor in educational discourse (Cameron 2003; Wan and Low 2015) to explore practice and attitude towards metalanguage varieties and their helpfulness for different learning facets.

Factor 3: Metaphor Continuation in Written Interaction

Interestingly, this factor explained the same amount of variance as MLP (Comprehension) in Spoken Dialogue but contained only two substantial loadings, Q12-Im completely_W_P (0.84) and Q11-magic circle commend him for_W_P (0.46). For Q12-Im completely_W_P, 35 per cent (24/69) of productions successfully continued the dialogue's figurative meaning, most focusing on the effect on the narrator: 'under his spell (P5)', 'surprised (P6)', 'enchanted (P10, P34, P36, P67)'. One deferential production referenced Harry Potter: 'a muggle comparing with him (P50)'; and a few productions conveyed supportiveness, for example, 'convinced that he deserves this magic award (P47)'.

Qualitatively, Q11-magic circle commend him for_W_P productions depicted work-related performance via specific spells: 'his flying magic spell? (P18)'; magical achievement, 'his magical achievement? (P3)'; performance, 'his great performance in the show? (P13)'; or fighting danger, 'defeating monsters (P34)'. Less successful productions, scoring 1/2, invoked vaguer figurative meanings: 'miracle (P12)'; a person/position rather than achievement, 'Harry Porter (P50)'; a metaphorical prize rather than achievement, 'the treasure of the kingdom (P60)'; or spells generally 'his spells (P4)'. A minority of partially successful productions conveyed a potentially figurative or literal meaning, as in 'the award? (P2)'.

In the 20 per cent (14/69) responses scoring 0, the most common reason for this was that no response was attempted. Where an attempt was made, unsuccessful productions were either indecipherable, outside, or counter to the scope of possible meanings, for example, 'the best policy? His honesty? (P37)' which, if a creative rehash of 'honesty is the best policy', would not fit the context; 'Improvement (P8)' which jars with the achievement theme; and 'a long time? (P24)', where the significance of commending someone for a long time is unclear.

Compared with O'Reilly (2017), the Q12-Im completely_W_P data show proportionately fewer successful or unsuccessful productions and more partially successful productions, with 35 per cent versus 43 per cent scoring 2, 45 per cent versus 13 per cent scoring 1, and 20 per cent versus 45 per cent scoring 0. Q11-magic circle commend him for_W_P had just 23 per cent (16/69) successful productions scoring 2/2, although the current study's participants outperformed those

in O'Reilly (2017), with 23 per cent versus 13 per cent scoring 2, 43 per cent versus 36 per cent scoring 1, and 33 per cent versus 52 per cent scoring 0. Taken together, these findings suggest that more advanced overall language proficiency seems to be accompanied by improvements at the bottom end, specifically, a shift from unsuccessful to partially successful Metaphor Continuation in Written Interaction. However, the general trajectory in this MLP area could be better understood via further, longitudinal (within-participants) research.

Limitations

As an investigation scoping elicited, rather than naturalistic, MLP across modes/knowledges, it is unclear whether findings would reliably predict the quantity/quality of MLP in everyday, non-research discourse. Similarly, the extent to which findings generalize from the MLP types studied (idiom re-literalization, metaphor continuation etc.) to others (e.g. puns, teasing) is unknown. Analysis of MLP-spoken-mode-P responses was restricted to scoring for overall quality and coding for linguistic/conceptual/metalinguistic characteristics, meaning that parallels with research on fluid intelligence dimensions such as the number, originality, and speed of metaphor productions (cf. Johnson and Rosano 1993; Littlemore 2001), although intriguing, are not possible.

A second area of limitation concerns the effectiveness of the spoken-mode MLP tests. As noted, MLP-spoken-mode-R required the use of subtitles (written input) alongside aural/visual input, and so the extent of participants' reliance on either input is unclear. MLP-spoken-mode-P responses were elicited in a relatively relaxed, online interview rather than in-person examination/test conditions (cf. Littlemore 2001), which was likely conducive to participants revealing their true MLP abilities and boundaries of knowledge (Low 1988). However, despite several rich productions, MLP-spoken-mode-P items had mixed success in drawing participants away from to-the-point, task-focused responses to more creative, extended MLP, a reminder that even elicitation approaches do not always perfectly tap the intended construct. In future, MLP researchers might harness the combined strengths of elicited and naturalistic approaches where time/resources allow (O'Reilly and Marsden 2021), for example, a naturally observed instance of learner MLP used as the basis for a conversation with further, researcher-led MLP input.

A third set of limitations concern methodology. Overall, the robustness of the Exploratory Factor Analysis in providing insights into the elicited MLP construct is attested to via the homogeneity of participants, controlled procedure, preliminary and post hoc adequacy, and comparability of variance explained with previous L2 Metaphoric Competence modelling (O'Reilly and Marsden 2021). Nevertheless, we recognize that our sample size would be located on the lower end of those synthesized from 51 L2 studies using Exploratory Factor Analysis (Plonsky and Gonulal 2015), and so encourage researchers extending the current enquiry to strive for larger pools of participants where practical, which would also enable modelling of a wider range of parameters (see Note 4). Although Principal Axis Factoring (used on account of multivariate nonnormality) typically generates a hypothesis for future testing but limits conclusions to the sample rather than offering generalizations to the population (cf. Maximum-Likelihood, see Field 2013), the bootstrap factor loadings offered some insight into the stability of the solution from an internal replicability perspective (Zientek and Thompson 2007). Nevertheless, further investigation involving external replication of various kinds (e.g. close, partial, approximate, conceptual, see Porte and McManus 2019) is key to understanding the generalizability of the current study's findings to different MLP and learner types, and the role of MLP within metaphoric and broader language competences (O'Reilly and Marsden 2023).

Another methodological issue concerns reactions to the gender-sensitive ice cream metaphor (Low 1988), which may have been shaped by the predominantly female sample and interviewer (female, the second author). Seemingly, the social sensitivity of the metaphor diminished when the target of the metaphor shifted from the participant themselves to their friends/family, but further research is needed to understand the complexities of how, when, and why such metaphors prompt adverse reactions, at both the group and individual level.

A final issue concerns the need to remove three receptive metaphor continuation variables (Qs4–6) and one MLP in spoken production variable (Q18) to improve correlation compactness and the likelihood of distinct, reliable factors (Field 2013). Hence, the Exploratory Factor Analysis and subsequent analysis of linguistic/conceptual/metalinguistic themes focused on most (16/21) but not all, of the measured MLP variables. While further enquiry is needed to understand whether the diminished factorability from these four variables is sample- or construct-specific, we encourage researchers to make such data available as a move towards better reporting in this area (Plonsky and Gonulal 2015), and to scrutinize and, where necessary, take steps to maximize data factorability and the informativeness of models.

Conclusion

The current study offers new insights into the L2 MLP construct for advanced L1 Mandarin L2 English learners via a combination of quantitative and qualitative approaches and drawing on previous research into (M)LP and Metaphoric Competence more generally. We have provided evidence of the relative ease/difficulty of different MLP tasks, the latent interrelatedness of MLP modes/knowledges, and a systematic characterization of the linguistic/conceptual/metalinguistic themes underpinning learners' MLP. While the findings may be of primary relevance to readers with an interest in the MLP construct, where relevant, potential pedagogical implications have been provided for those interested in MLP in ELT, who may wish to adapt some of the items to classroom tasks and/or obtain an approximate measure of their learners' MLP abilities. However, as a nonintervention study, where implications for pedagogy are discussed, these must remain tentative. Going forward, we hope that the substantive findings, availability of data/materials, and teaching implications assist researchers, educators, and learners to continue taking so-called non-serious language more seriously (Cekaite and Aronsson 2005).

Notes on contributors

David O'Reilly is a Lecturer in Education (University of York). His research investigates L1/L2 metaphoric competence, vocabulary knowledge, language testing/development, and language play. David was the RaAM Early Career Research Paper Prize 2021 winner for his article 'Eliciting and measuring L2 metaphoric competence: Three decades on from Low (1988)' published in *Applied Linguistics*. He is involved in research projects on L2 grammar processing and promoting Open Research and is *Metaphor and the Social World* Review Editor. *Address for correspondence*: Centre for Advanced Studies in Language and Education, Department of Education, University of York, York YO10 5DD, UK. <david.oreilly@york.ac.uk>

Luling Yan is a Mandarin Immersion Class Teacher in Kensington Elementary School (North Carolina, USA). Having attained an MA in Applied Linguistics for English Language Teaching (University of York, 2019), she is completing Teach Now and Foundations of Mathematics teacher training programmes at Morland University. Luling has presented at the Chinese Language Teachers Association conference and hosted workshops for teachers on student-centred teaching strategies. She was a 2023 Participate Learning Teacher of the Year candidate.

Notes

[1] Participants had been learning English for 10–18 years and had passed TEM-8 in the years 2019–2013 (n = 9, 19, 3, 23, 6, 5, 2, respectively) and 2010 (n = 2), as a group averaging a high 'pass' score (M = 68.01, SD = 7.18), n = 50 scoring 60–69/'pass', n = 13 scoring 70–79/'good', n = 6 scoring ≥80/'excellent'.

[2] In O'Reilly and Marsden (2021), Test 9_Metaphor continuation-R did not strongly load on the latent MLP factor. However, since it was devised/refined to tap comprehension of a playful aspect of metaphor use (Low 1988; O'Reilly 2017; O'Reilly and Marsden 2021) in the current study, it was selected as a natural counterpart to the corresponding productive test.

[3] We introduced one amendment to O'Reilly and Marsden's (2021) metaphor continuation production scoring rubric, allowing potentially literal responses to score 1 (partially successful) rather than harshly restricting them to 0 (unsuccessful).
[4] As an early endeavour in L2 MLP modelling, the current study employed Exploratory Factor Analysis as a straightforward and familiar way of exploring multidimensionality, allowing for comparisons with O'Reilly and Marsden (2021), and laying the foundation for further MLP construct research and studies with an MLP test refinement focus. The use of bootstrapping offered an indication of the stability/internal replicability of the various item-factor loadings. In future research, where study design allows, we would advocate more advanced approaches such as exploratory and confirmatory multidimensional IRT (Reckase 2009) in favour of those grounded in classic test theory, since these offer a highly flexible and robust means of investigating dimensionality and various test item properties (e.g. difficulty, discriminability), although stable parameter estimation would require larger samples sizes than were possible in the current study (Immekus *et al.* 2019).
[5] For readers interested in cross-referencing these examples with participants' performance on the different MLP factors, Supplementary Table S3 provides factor scores calculated using Thurstone's (1935) regression method to permit possible interrelatedness (Field 2009), with corresponding participant ranks.
[6] Two items had negative 95 per cent confidence interval lower bounds: Factor 2—Q13_main_meaning_S_R_fig (0.36 [–0.01, 0.73]), and factor 3—Q11_magic_circule_commend_him_for_W_P_fig (0.46 [–0.11, 1.03]), but since these were close to 0 and the majority of resamples were positive, these loadings were also deemed to be relatively stable.
[7] We are grateful to an anonymous reviewer for this suggestion and the 'bird's eye view' metaphor.

References

Attardo, S. 1994. *Linguistic Theories of Humour*. Mouton de Gruyter. https://doi.org/10.1515/9783110219029

Attardo, S. 2001. *Humorous Texts: A Semantic and Pragmatic Analysis*. Mouton de Gruyter. https://doi.org/10.1515/9783110887969

Attardo, S. and V. Raskin. 1991. 'Script theory revis(it)ed: Joke similarity and joke representational model,' *Humor: International Journal of Humor Research* 4/3: 293–347. https://doi.org/10.1515/humr.1991.4.3-4.293.

Bakhtin, M. [1934] 1981. *The Dialogic Imagination, Four Essays by M. M. Bakhtin* in M. Holoquist (ed.), translated by C. Emerson and M. Holoquist. University of Texas Press.

Bates, E., *et al.* 1980. 'The role of pronominalization and ellipsis in texts: Some memory experiments,' *Journal of Experimental Psychology: Human Learning and Memory* 6: 676–91. https://doi.org/10.1037/0278-7393.6.6.676.

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. https://doi.org/10.1093/applin/amh043.

Bell, N. 2012a. 'Comparing playful and nonplayful incidental attention to form,' *Language Learning* 62/1: 236–65. https://doi.org/10.1111/j.1467-9922.2011.00630.x.

Bell, N. 2012b. 'Formulaic language, creativity, and language play in a second language,' *Annual Review of Applied Linguistics* 32: 189–205. https://doi.org/10.1017/s0267190512000013.

Belz, J. A. 2002. 'The myth of the deficient communicator,' *Language Teaching Research* 6/1: 59–82. https://doi.org/10.1191/1362168802lr097oa.

Belz, J. A. 2009. 'Second language play as a representation of the multicompetent self in foreign language study,' *Journal of Language, Identity, and Education* 1/1: 13–39. https://doi.org/10.1207/S15327701JLIE0101_3.

Boers, F. 2000. 'Metaphor awareness and vocabulary retention,' *Applied Linguistics* 21/4: 553–71. https://doi.org/10.1093/applin/21.4.553.

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. https://doi.org/10.1177/1362168806072460.

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

Broner, M. A. and E. E. Tarone. 2001. 'Is it fun? Language play in a fifth-grade Spanish immersion classroom,' *The Modern Language Journal* 85/3: 363–79. https://doi.org/10.1111/0026-7902.00114.

Bushnell, C. 2009. '"Lego my keego!": An analysis of language play in a beginning Japanese as a foreign language classroom,' *Applied Linguistics* 30/1: 49–69. https://doi.org/10.1093/applin/amn033.

Cameron, L. 2003. 'Metaphor in educational discourse,' *Continuum* https://doi.org/10.5040/9781474212076.

Carroll, J. B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Cekaite, A. and K. Aronsson. 2005. 'Language play, a collaborative resource in children's L2 learning,' *Applied Linguistics* 26/2: 169–91. https://doi.org/10.1093/applin/amh042.

Chen, W.-H., McLeod, L. D. and T. M. Coles. 2014a. 'Rasch First? Factor First?' Presented at ISPOR 17th Annual European Congress. Nov 8-12 (2014), Amsterdam, The Netherlands.

Chen, W.-H., L. D. McLeod, and T. M. Coles. 2014b. 'Rasch first? Factor first?,' *Value in Health* 17/7: A569. https://doi.org/10.1016/j.jval.2014.08.1901.

Christensen, K. B., G. Engelhard Jr, and T. Salzberger. 2012. 'Ask the experts: Rasch vs Factor Analysis,' *Rasch Measurement Transactions* 26/3: 1373–8.

Cohen, L., L. Manion, and K. Morrison. 2017. *Research Methods in Education*. 8th edn.. Routledge. https://doi.org/10.4324/9781315456539

Cook, G. 1997. 'Language play, language learning,' *ELT Journal* 51/3: 224–31. https://doi.org/10.1093/elt/51.3.224.

Cook, G. 2000. *Language Play, Language Learning*. Oxford University Press.

Craik, F. I. M. and R. Lockhart. 1972. 'Levels of processing: A framework for memory research,' *Journal of Verbal Learning and Verbal Behavior* 11: 671–84. https://doi.org/10.1016/S0022-5371(72)80001-X.

Crystal, D. 1996. 'Language play and linguistic intervention,' *Child Language Teaching and Therapy* 12/3: 328–44. https://doi.org/10.1177/026565909601200307.

Crystal, D. 1998. *Language Play*. Penguin.

Dore, M. 2015. 'Metaphor, humour and characterisation in the TV comedy programme Friends,' *Cognitive Linguistics and Humor Research* 26/9: 191–214. https://doi.org/10.1515/9783110346343-010.

Eisinga, R., M. T. Grotenhuis, and B. Pelzer. 2013. 'The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?,' *International Journal of Public Health* 58: 637–42. https://doi.org/10.1007/s00038-012-0416-3.

Fabrigar, L. R., *et al*. 1999. 'Evaluating the use of exploratory factor analysis in psychological research,' *Psychological Methods* 4/3: 272–99. https://doi.org/10.1037/1082-989X.4.3.272

Fauconnier, G. and M. Turner. 1996. 'Blending as a central process of grammar' in , A. Goldberg (ed.): *Conceptual Structure, Discourse, and Language*. Cambridge University Press, pp. 113–29.

Fauconnier, G. and M. Turner. 1998. 'Conceptual integration network,' *Cognitive Science* 22/2: 133–87. https://doi.org/10.1207/s15516709cog2202_1.

Fauconnier, G. and M. Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

Field, A. 2013. *Discovering Statistics Using IBM SPSS Statistics*. 4th edn. Sage.

Immekus, J. C., K. E. Snyder, and P. A. Ralstdon. 2019. 'Multidimensional item response theory for factor structure assessment in educational psychology research,' *Frontiers in Education* 4/45: 1–15. https://doi.org/10.3389/feduc.2019.00045.

Jin, Y. and J. Fan. 2011. 'Test for English Majors (TEM) in China,' *Language Testing* 28/4: 589–96. https://doi.org/10.1177/0265532211414852.

Johnson, J. and T. Rosano. 1993. 'Relation of cognitive style to metaphor interpretation and second language proficiency,' *Applied Psycholinguistics* 14/2: 159–75. https://www.doi:10.1017/S014271640000953X.

Knoch, U. and McNamara, T. 2015. 'Rasch analysis' in L. Plonsky (ed.): *Advancing Quantitative Methods in Second Language Research*. Routledge pp. 275–304. https://doi.org/10.4324/9781315870908

Lakoff, G. and M. Johnson. 1980. *Metaphors We Live by*. University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226470993.001.0001

Landis, J. R. and G. G. Koch. 1977. 'The measurement of observer agreement for categorical data,' *Biometrics* 33: 159–74. https://doi.org/10.2307/2529310.

Lantolf, J. P. 1997. 'The function of language play in the acquisition of L2 Spanish' in W. R. Glass and A. T. Perez-Leroux (eds): *Contemporary Perspectives on the Acquisition of Spanish*. Cascadilla Press, pp. 3–24.

Laufer, B. and Z. Goldstein. 2004. 'Testing vocabulary knowledge: Size, strength, and computer adaptiveness,' *Language Learning* 54/3: 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x.

Littlemore, J. 2001. 'Metaphoric competence: A language learning strength of students with a holistic cognitive style?,' *TESOL Quarterly* 35: 459–91. https://doi.org/10.2307/3588031.

Littlemore, J. 2008. 'The relationship between associative thinking, analogical reasoning, image formation and metaphoric extension strategies' in M. S. Zanotto, L. Cameron, and M. C. Cavalcanti (eds.): *Confronting Metaphor in Use: An Applied Linguistic Approach*. John Benjamins, pp. 199–222. https://doi.org/10.1075/pbns.173.14lit

Littlemore, J. and G. D. Low. 2006. *Figurative Thinking and Foreign Language Learning*. Palgrave Macmillan. https://doi.org/10.1057/9780230627567

Loewen, S. and T. Gonulal. 2015. 'Exploratory factor analysis and principal components analysis' in L. Plonsky (ed.): *Advancing Quantitative Methods in Second Language Research*. Routledge, pp. 182–212. https://doi.org/10.4324/9781315870908

Low, G. D. 1988. 'On teaching metaphor,' *Applied Linguistics* 9: 125–47. https://doi.org/10.1093/applin/9.2.125.

Lucas, T. 2005. 'Language awareness and comprehension through puns among ESL learners,' *Language Awareness* 14: 221–38. https://doi.org/10.1080/09658410508668838.

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, pp. 155–73. https://doi.org/10.1075/aila.23.09mac

McDaniel, M., *et al*. 1995. 'The bizarreness effect: It's not surprising, it's complex', *Journal of Experimental Psychology: Learning, Memory and Cognition* 21: 422–35. https://doi.org/10.1037/0278-7393.21.2.422.

McNeish, D. 2018. 'Thanks coefficient alpha, we'll take it from here,' *Psychological Methods* 23: 412–33. https://doi.org/10.1037/met0000144.

Miller, N. 2011. 'The processing of malformed formulaic language,' *Applied Linguistics* 32/2: 129–48. https://doi.org/10.1093/applin/amq035.

Murakami, A. and T. Alexopoulou. 2016. 'L1 influence on the acquisition order of English grammatical morphemes,' *Studies in Second Language Acquisition* 38: 365–401. https://doi.org/10.1017/s0272263115000352.

Nacey, S. 2013. *Metaphors in Learner English*. John Benjamins. https://doi.org/10.1075/milcc.2

Naciscione, A. 2020. 'Reproducibility of patterns of stylistic use of phraseological units: A cognitive diachronic view' in M. Omazić and J. Parizoska (eds): *Reproducibility and Variation of Figurative Expressions: Theoretical Aspects and Applications*. University of Białystok Publishing House, pp. 33–50.

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. 2021. 'Eliciting and measuring L2 metaphoric competence: Three decades on from Low (1988),' *Applied Linguistics* 42/1: 24–59. https://doi.org/10.1093/applin/amz066.

O'Reilly, D. and E. Marsden. 2023. 'Elicited metaphoric competence in a second language: A construct associated with vocabulary knowledge and general proficiency?,' *International Review of Applied Linguistics in Language Teaching* 61/2: 287–327. https://doi.org/10.1515/iral-2020-0054.

Philip, G. 2010. '"Drugs, traffic, and many other dirty interests": Metaphor and the language learner' in G. Low, Z. Todd, A. Deignan, and L. Cameron (eds): *Researching and Applying Metaphor in the Real World*. John Benjamins, pp. 63–80. https://doi.org/10.1075/hcp.26.05phi

Pitzl, M.-L. 2016. 'World Englishes and creative idioms in English as a lingua franca,' *World Englishes* 35: 293–309. https://doi.org/10.1111/weng.12196.

Plonsky, L. and D. Derrick. 2016. 'A meta-analysis of reliability coefficients in second language research,' *The Modern Language Journal* 100: 538–53. https://doi.org/10.1111/modl.12335.

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. https://doi.org/10.1111/lang.12111.

Porte, G., and K. McManus. 2019. 'What kind of replication should you do?' in G. Porte, and K. McManus (eds): *Doing Replication Research in Applied Linguistics*. Routledge, pp. 69–94. https://doi.org/10.4324/9781315621395

Prodromou, L. 2003. 'The idiomatic paradox and English as a lingua franca,' *Modern English Teacher* 12/1: 22–9.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available at https://www.R-project.org/. Accessed 13 April 2023.

Read, J. 1998. 'Validating a test to measure depth of vocabulary knowledge' in A. Kunnan, (ed.): *Validation in Language Assessment*. Lawrence Erlbaum, pp. 41–60. https://doi.org/10.4324/9780203053768

Reckase, M. D. 2009. *Multidimensional Item Response Theory*. Springer. https://link.springer.com/book/10.1007/978-0-387-89976-3

Schmidt, S. and A. Williams. 2001. 'Memory for humorous cartoons,' *Memory and Cognition* 29: 305–11. https://doi.org/10.3758/BF03194924.

Smith, R. 1996. 'A comparison of methods for determining dimensionality in Rasch measurement,' *Structural Equation Modeling – A Multidisciplinary Journal* 3/1: 25–40. https://doi.org/10.1080/10705519609540027.

Strick, M., *et al*. 2010. 'Humor in the eye tracker: Attention capture and distraction from context cues,' *Journal of General Psychology* 137: 37–48. https://doi.org/10.1080/00221300903293055.

Sullivan, P. 2000. 'Playfulness as mediation in communicative language teaching in a Vietnamese classroom' in J. P. Lantolf (ed.): *Sociocultural Theory and Second Language Learning*. Oxford University Press, pp. 115–31.

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

Tarone, E. 2000. 'Getting serious about language play: Language play, interlanguage variation and second language acquisition' in B. Swierzbin *et al*. (eds): *Social and Cognitive Factors in Second Language Acquisiton*. Sommerville, MA: Cascadilla Press, pp. 31–54.

Tarone, E. 2005. 'Fossilization, social context, and language play' in Z. Han and T. Odlin (eds): *Studies of Fossilization in Second Language Acquisition*. Multilingual Matters, pp. 157–72. https://doi.org/10.21832/9781853598371-010

Thurstone, L. L. 1935. *The Vectors of Mind*. University of Chicago Press.

Tocalli-Beller, A. and M. Swain. 2007. 'Riddles and puns in the ESL classroom: Adults talk to learn' in A. Mackey (ed.): *Conversational Interaction in Second Language Acquisition: Empirical Studies*. Oxford University Press, pp. 143–67.

Trahey, M. and L. White. 1993. 'Positive evidence and pre-emption in the second language classroom,' *Studies in Second Language Acquisition* 15: 181–204. https://doi:10.1017/S0272263100011955.

Wan, W. and G. D. Low. 2015. *Elicited Metaphor Analysis in Educational Discourse*. John Benjamins. https://doi.org/10.1075/milcc.3

Webb, S. and M. P. H. Rodgers. 2009a. 'The lexical coverage of movies,' *Applied Linguistics* 30/3: 407–27. https://doi.org/10.1093/applin/amp010.

Webb, S. and M. P. H. Rodgers. 2009b. 'The vocabulary demands of television programs,' *Language Learning* 59/2: 335–66. https://doi.org/10.1111/j.1467-9922.2009.00509.x.

Werkmann Horvat, A., *et al*. 2022. 'Comprehension of different types of novel metaphors in monolinguals and multilinguals,' *Language and Cognition* 14/3: 509–509. https://doi:10.1017/langcog.2022.8.

Yan, L. 2019. '*Playing with metaphor in elicited speaking and writing: The same in a first and second language? The case of Chinese EFL learners*', MA thesis, University of York.

Zientek, L. R. and B. Thompson. 2007. 'Applying the bootstrap to the multivariate case: Bootstrap component/factor analysis,' *Behavior Research Methods* 39: 318–25. https://doi.org/10.3758/BF03193163.