

UC Santa Barbara

UC Santa Barbara Previously Published Works

Title

Individual differences in speakers' perceptions of psycholinguistic dimensions: Modeling idiom processing advantage in a phrase judgment task among L1 and L2 speakers

Permalink

<https://escholarship.org/uc/item/3cz723kz>

Journal

Bilingualism Language and Cognition

ISSN

1366-7289

Authors

Zhong, Yanlu
Brehm, Laurel

Publication Date

2026-07-15

DOI

10.1017/s1366728926101606

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Running-head: Individual Differences in Idiom Processing

Individual Differences in Speakers' Perceptions of Psycholinguistic Dimensions: Modeling Idiom Processing Advantage in a Phrase Judgment Task among L1 and L2 Speakers

Yanlu Zhong
University of California, Santa Barbara

Laurel Brehm
University of California, Santa Barbara

*Acknowledgments: The authors thank the editor and the anonymous reviewers for their constructive comments and helpful suggestions.

Competing interests: The authors declare none.

Address for correspondence:
Yanlu Zhong
E-mail: yanlu_zhong@ucsb.edu
Mailing address: UC Santa Barbara, Santa Barbara, CA 93106

Keywords: individual differences; idiom processing; phrase judgment task; psycholinguistic dimensions; L2 speakers

Abstract

This study investigates whether individual differences (IDs) in speakers' perceptions of psycholinguistic dimensions (familiarity, knowledge, transparency, ambiguity, valence, arousal) predict idiom processing advantage over matched novel control phrases in L1 and L2 speakers in a phrase judgment task. Bayesian multilevel models showed that item-level average norms from L1 speakers for familiarity, valence, and arousal facilitated idiom judgment in both groups, whereas transparency and ambiguity had inhibitory effects. Beyond item means, deviations in speaker-specific perceptions from the item-level average norm predicted idiom judgment advantage. In both groups, individuals who rated an idiom as more positive and more ambiguous than its item-level average norm showed larger idiom judgment advantages. In L1-L2 comparisons, individual-level deviation effects of valence and transparency were stronger in L2 than in L1 speakers, whereas ambiguity deviation effects were stronger in L1 speakers. Overall, modeling ID effects beyond group norms reveals finer-grained L1-L2 differences in language processing.

Keywords: individual differences; idiom processing; phrase judgment task; psycholinguistic dimension; L2 speakers

Introduction

Individuals differ in how they engage with structural, semantic, and emotional language features (Kidd et al., 2018). Idioms, whose meanings are not fully inferable from literal components (Gibbs, 1994), provide a useful lens on IDs in language processing. Prior work shows that idioms are often processed faster and more accurately than comparable literal phrases, especially by L1 speakers. This idiom processing advantage over matched novel control phrases has been linked to psycholinguistic dimensions such as familiarity and ambiguity (Libben & Titone, 2008). However, idioms also carry affective meaning as part of their figurative interpretations, making valence and arousal relevant dimensions for modeling idiom processing (Citron et al., 2016; Morid & Sabourin, 2023, 2024). Most studies use L1 group-level ratings to provide standardized, stable, and comparable indices of these properties (Siyanova-Chanturia et al., 2011; Milburn et al., 2021; Soto-Sierra & Ferreira, 2024). Although useful, such norms may not fully capture IDs among L1 speakers or the experiential and cognitive realities of L2 learners, whose processing variability is well documented (Ellis, 2004). To address these questions, L1 and L2 participants completed a phrase judgment task and a subjective ratings task. This study analyzed six psycholinguistic dimensions, namely familiarity, knowledge, transparency, ambiguity, valence, and arousal, at two levels: (1) Item-level norms were estimated from L1 group-level mean ratings because L1 ratings are commonly used as target-language norms in idiom research (e.g., Bulkes & Tanner, 2017; Libben & Titone, 2008; Titone & Connine, 1994). These norms provided a common baseline for comparing L1 and L2 speaker-specific deviations and were used to approximate relatively stable properties of the expression. (2) Speaker-specific ratings

captured each individual's perception or experience of the same idiom. We then examined whether speaker-specific deviations from item-level average ratings (deviation scores) predicted the idiom judgment advantage, defined as faster responses to idioms than matched novel control phrases in the meaningfulness judgment task. The findings refine theoretical accounts of idiom processing by showing that IDs across psycholinguistic dimensions are integral to how idioms are represented and accessed, rather than mere noise.

Literature Review

Idiom processing in L1 and L2 speakers

Idioms, such as *spill the beans*, are expressions whose figurative meanings are not always inferable from their constituent words (Gibbs, 1994). The dual-route model (Titone & Connine, 1999) characterizes idiom processing as the joint outcome of two parallel, interacting mechanisms: a holistic route, which retrieves idiomatic meaning from stored multiword representations, and a compositional route, which incrementally constructs meaning from constituent words. The relative contribution of these routes is sensitive to psycholinguistic dimensions such as familiarity (Siyanova-Chanturia et al., 2011).

Comparisons between L1 and L2 speakers suggest that the two routes are weighted differently across groups. L1 speakers tend to engage holistic retrieval together with compositional analysis, allowing relatively rapid access to idiomatic meanings. Consistent with this account, L1 speakers show an idiom processing advantage over matched literal expressions, including faster processing in minimal contexts (Carrol & Conklin, 2014) and shorter and fewer fixations on idioms than on literal controls (Siyanova-Chanturia et al., 2011). However, L1 processing is not exclusively holistic: literal interpretations may be

activated early and delay figurative access when an idiom becomes identifiable only late in the string (Cacciari & Tabossi, 1988). Eye-tracking evidence likewise suggests simultaneous recruitment of both routes, with their relative weighting shaped by how strongly component words support the figurative meaning (Titone & Connine, 1999). In L2 speakers, by contrast, the compositional route may remain more strongly weighted than in L1 processing, even among highly proficient L2 speakers, given their less naturalistic pre-immersion English learning or use and comparatively less cumulative idiom-specific exposure than L1 speakers. Studies of L2 English speakers suggest that idiom-processing advantages are often weaker or less consistent than in L1 speakers (Siyanova-Chanturia et al., 2011). Evidence from highly proficient German-Dutch bilinguals with several years of residence in the L2 environment further shows activation of both figurative and literal meanings during L2 idiom processing, while sensitivity to idiom properties remains reduced relative to L1 speakers (van Ginkel & Dijkstra, 2020). Together, these findings support a route-weighting account in which L1 and L2 speakers recruit both pathways but differ in their default balance and in the conditions enabling reliable holistic retrieval.

The dual-route framework, therefore, accommodates ID effects: the balance between holistic retrieval and compositional analysis should vary with both item-level psycholinguistic properties and speakers' experience and knowledge. In L1 speakers, such variability likely reflects differences in the efficiency of holistic access and the use of compositional cues across idioms. In L2 speakers, for whom processing more often relies on compositional analysis and holistic access depends more on supportive cues, IDs may be larger and more consequential, reflecting greater heterogeneity in experience, proficiency,

and representational stability. This motivates testing whether item-level average norms and speaker-specific perceptions moderate route weighting similarly across L1 and L2 speakers, and whether ID effects capture systematic cue-weighting differences rather than residual noise.

Psycholinguistic dimensions in idiom processing

Idiom processing is shaped by multiple psycholinguistic dimensions, including familiarity, knowledge, transparency, ambiguity, valence, and arousal (e.g., Morid & Sabourin, 2023). In prior work, these dimensions have typically been operationalized as item-level average normed ratings from L1 speakers, although they may obscure speaker-specific perception variation.

Familiarity indexes prior exposure and recognition (Abel, 2003) and robustly facilitates idiom recognition and interpretation in L1 speakers, although effects may attenuate when familiarity is near ceiling (e.g., Carrol & Conklin, 2014). In L2 processing, familiarity can still support figurative activation, but its impact is often less differentiated, consistent with restricted exposure and reduced variability in familiarity within L2 samples (Conklin & Schmitt, 2008; Titone et al., 2015).

Knowledge can refer to objectively verified knowledge of an idiom's conventional figurative meaning or to speakers' self-rated access to that meaning. The present study focuses on the latter, as perceived knowledge may influence whether an idiom is treated as meaningful and readily available during the judgment task, although it does not verify objective meaning accuracy. In L1 speakers, knowledge supports rapid figurative retrieval via entrenched lexicalized representations (Abel, 2003). In L2 speakers, especially those with

less entrenched idiom representations, knowledge may play a comparatively larger role because idiom comprehension depends on the stability of associations between an idiom's phrase form and its figurative meaning (e.g., O'Reilly et al., 2025). Knowledge also covaries with other dimensions: in L1 speakers, familiarity reliably predicts idiom knowledge, whereas in L2 speakers, cross-language similarity is a stronger predictor (Titone & Connine, 1994; Soto-Sierra & Ferreira, 2024).

Transparency refers to the extent to which figurative meaning can be inferred from constituents (Siyanova-Chanturia et al., 2011). Findings for L1 speakers are mixed: transparency effects can be weak when idioms are highly familiar, yet more transparent idioms are sometimes processed more readily (e.g., Swinney & Cutler, 1979; Titone & Connine, 1999). In L2 processing, transparency may be especially relevant because learners often rely more strongly on analytic, word-based routes (Steinel et al., 2007). This aligns with dual-route accounts in which analytic and holistic mechanisms jointly contribute, with relative weighting varying across speaker groups and item properties (Libben & Titone, 2008).

Ambiguity captures competition between literal and figurative interpretations (Titone & Connine, 1994). In L1 speakers, ambiguity can facilitate processing in literal-biased contexts, whereas in L2 speakers it more often increases difficulty, given stronger initial reliance on literal meanings and delayed figurative activation (Cronk & Schweigert, 1992; Siyanova-Chanturia et al., 2011; Milburn & Warren, 2019). Importantly, effects depend on context and proficiency, and divergent patterns among advanced L2 learners suggest meaningful individual variability in how ambiguity is resolved (Milburn et al., 2021).

Idioms also carry emotional meaning as part of their figurative interpretations, and this affective content may influence how they are represented and processed. Valence and arousal, defined as pleasantness and physiological activation (Russell, 2003), vary systematically across idioms, showing that emotional meaning characterizes figurative phrases as well as individual words (Citron et al., 2016; Morid & Sabourin, 2023, 2024). These affective dimensions influence word-level processing alongside established psycholinguistic dimensions such as ambiguity, concreteness, and imageability, providing a basis for examining their role in idiom processing (Citron, 2012; Kuperman et al., 2014; Morid & Sabourin, 2023, 2024). These affective dimensions are especially relevant for L1-L2 comparisons because emotional effects in L2 speakers may be weaker or more variable, depending on learning context and language experience (Pavlenko, 2012; Sheikh & Titone, 2016). Recent work further suggests that valence and arousal influence idiom processing and interact with core psycholinguistic properties (Morid & Sabourin, 2023). By including valence and arousal alongside familiarity, knowledge, transparency, and ambiguity, the present study tests whether affective meaning provides processing-relevant information beyond form-based and semantic-inferential cues, and whether these effects differ across language backgrounds.

IDs in speakers' perceptions of psycholinguistic dimensions

IDs refer to systematic variation across speakers in cognitive, linguistic, and experiential factors that shape how language is perceived, represented, and processed (Kidd et al., 2018).

In L1 speakers, IDs systematically moderate language processing across

representational levels. For example, IDs in spelling proficiency were found to modulate neighborhood-priming effects in L1 word recognition (Andrews & Hersch, 2010), and IDs in lexical familiarity have been shown to shape performance in naming, lexical decision, and semantic categorization tasks (Lewellen et al., 1993). At the phrasal level, linguistic knowledge and processing skills modulate contextual facilitation in idiom processing (Tilmatine et al., 2021), while idiom comprehension also varies with working memory, inhibitory control, and personality-related factors (Cacciari et al., 2018). Together, these findings show that L1 ID effects extend beyond domain-general cognitive abilities to include speaker-specific linguistic experience.

In L2 speakers, IDs may be especially consequential because processing is shaped by cognitive capacities, proficiency, immersion and language-use history. At the word level, IDs in lexical entrenchment shape L2 word recognition, with less entrenched representations showing stronger frequency effects (Diependaele et al., 2013). Similarly, IDs in L2 experience modulate frequency effects during reading, suggesting that language-use history affects lexical access (Whitford & Titone, 2012). Yi (2018) showed that advanced L2 speakers' processing of adjective-noun collocations varies with cognitive profiles linked to implicit and explicit aptitude. Similar ID-driven variation occurs in idiom processing: even in literal-biased contexts, some L2 learners activate figurative meanings, suggesting that idiom conventionality can outweigh contextual information for certain individuals (Milburn et al., 2021). Beyond overgeneralized figurative activation, some L2 speakers may rely more on compositional analysis. Milburn et al. (2026) found that L1 and advanced L2 speakers use different idiom-processing strategies, with L2 speakers showing greater compositional

reliance. Such variability in L2 idiom processing provides a useful test case for whether speaker-specific ratings capture variation not fully represented by item-level average norms.

A key limitation of norm-based idiom research is that item-level norms may obscure speaker-specific variation in how idioms are perceived and processed. Rating tasks are commonly used to assess psycholinguistic dimensions, and most studies average L1 ratings to create normed values for use in both L1 and L2 research (Milburn et al., 2021; Siyanova-Chanturia et al., 2011; Soto-Sierra & Ferreira, 2024). This practice assumes that L1 speakers provide stable, low-noise estimates of key properties, yielding reliable baselines for cross-group comparisons. However, aggregate norms may obscure meaningful individual variation: two speakers may rate the same idiom very differently, producing processing differences that disappear when ratings are averaged. Incorporating speaker-specific ratings, therefore, provides a more ecologically valid and precise account of how mental representations relate to idiom processing.

In the present study, each psycholinguistic dimension was analyzed at two distinct levels: item-level norms and speaker-specific ratings. Item-level norms, estimated from group mean ratings from L1 speakers, approximate relatively stable properties of each expression, whereas speaker-specific ratings capture individual perceptions or experiences of the same idiom. Although their relative importance may differ by dimension, both levels remain relevant: familiarity and knowledge are experience-based but can yield useful speech-community norms, whereas transparency may be more item-inherent but still varies in perceived form-meaning relations across speakers (e.g., Bulkes & Tanner, 2017; Libben & Titone, 2008; Titone & Connine, 1994; Citron et al., 2016; Morid & Sabourin, 2024).

L1-L2 Differences in ID effects

Usage-based accounts attribute L1-L2 differences to input history, with repeated, contextually rich experience strengthening form-meaning mappings. Richer exposure supports entrenched, semantically integrated representations, whereas limited input yields weaker, more form-driven representations (Bybee, 2010; Ellis, 2002). Accordingly, L1 idiom processing is supported by experience-based representations built through rich contextual exposure (Siyanova-Chanturia & Martinez, 2015). In contrast, L2 speakers often show qualitatively different processing profiles, even when they are highly proficient and immersed. The Representation Deficit Hypothesis proposes that L2 idioms may be stored as loosely connected lexical strings with reduced semantic integration relative to L1 representations (Abel, 2003). This account remains relevant to the present highly proficient L2 speakers with later immersion experience because target-language residence does not necessarily yield L1-like idiom comprehension (van Ginkel & Dijkstra, 2020; Suñer, 2019). The Dual Route Model further predicts that L1 speakers typically rely more on holistic processing for familiar idioms, whereas L2 speakers may rely more on analytic or literal-first parsing (Libben & Titone, 2008). This may also apply to highly proficient L2 speakers, reflecting less naturalistic pre-immersion English learning or use, weaker idiom-specific entrenchment, and less culturally embedded experience rather than limited general proficiency. These accounts suggest that idiom processing reflects cognitive capacities, experiential histories, and task demands, with systematic differences between L1 and L2 speakers.

Prior empirical research shows clear L1-L2 contrasts in idiom processing. L1 speakers

show a robust idiom processing advantage over novel control phrases, supported by rapid retrieval from entrenched representations (Swinney & Cutler, 1979; Carrol & Conklin, 2014; Siyanova-Chanturia et al., 2011). In contrast, L2 learners rely more on compositional, word-by-word analysis and show delayed or reduced figurative activation, yielding a smaller processing advantage (Conklin & Schmitt, 2008; Kyriacou & Köder, 2024; Senaldi et al., 2022). As reviewed above, the effects of item-level average norms for psycholinguistic dimensions diverge systematically between L1 and L2 speakers. Item-level average familiarity norms tend to facilitate figurative activation less robustly in L2, reflecting the critical role of reduced exposure in shaping idiom retrieval (Titone et al., 2015; Soto-Sierra & Ferreira, 2024). In contrast, item-level average knowledge norms may exert a disproportionately larger influence on L2 than on L1 speakers, consistent with evidence that L2 idiom representations may be less lexicalized or entrenched than L1 representations across both non-immersed and highly experienced L2 populations (Soto-Sierra & Ferreira, 2024; van Ginkel & Dijkstra, 2020; Suñer, 2019). Likewise, idioms with higher item-level average transparency norms are often easier for L2 speakers to interpret, whereas transparency effects in L1 speakers are more variable and task-dependent (Soto-Sierra & Ferreira, 2024). In addition, item-level average ambiguity and emotionality norms may impose greater cognitive costs in L2 than in L1 speakers, because resolving literal-figurative competition and emotional meaning can be more effortful without deeply entrenched conceptual grounding (Carrol & Conklin, 2017; Milburn & Warren, 2019).

IDs further accentuate L1-L2 contrasts. In L1 speakers, group-level ratings often provide reliable indices of psycholinguistic dimensions, with individual ratings adding finer-

grained variation (Ellis, 2006). In L2 speakers, however, IDs tend to exert stronger effects on online processing. For example, Yi (2018) found that L1 and advanced L2 speakers' adjective-noun collocation processing differed in relation to cognitive profiles, with L2 performance more strongly tied to implicit and explicit aptitude. Similarly, Morid and Sabourin (2024) reported greater L2 than L1 variability in speaker-specific ratings of idiom valence, arousal, familiarity, imageability, and concreteness. Thus, L1 speakers may rely on more stable, automatized idiom representations, whereas L2 speakers may rely on more variable learning experiences and therefore show stronger ID effects in perceptions of psycholinguistic dimensions.

The Current Study

Despite growing interest in idiom processing, several issues remain underexplored. First, most studies rely on item-level L1 ratings without testing whether IDs in speakers' perceptions of psycholinguistic variables better predict idiom processing advantage in real-time tasks such as phrase judgment tasks. Second, research has focused mainly on semantic properties such as familiarity and transparency, whereas emotional dimensions such as valence and arousal have received less attention despite their role in cognitive processing. Third, few studies have examined whether ID effects vary in scope or direction across L1 and L2 speakers. To address these gaps, the present study examined whether speaker-specific deviations from item-level average ratings across psycholinguistic dimensions predicted idiom judgment advantage in a phrase judgment task. This advantage was defined as faster meaningfulness-judgment processing for idioms than for matched novel control phrases.

RQ1: To what extent do deviations in speaker-specific perceptions from item-level

norms improve the overall prediction of idiom judgment advantage in L1 speakers and highly proficient L2 speakers with later immersion experience, and does this overall contribution differ across the two groups?

Support for RQ₁ would be shown by deviation-inclusive models outperforming norm-only models, as indicated by improved ELPD/LOOIC and Bayesian R². Group differences in this improvement would indicate that speaker-specific ratings contribute differently to prediction in L1 and L2 speakers.

RQ₂: How do deviations in speaker-specific perceptions of each psycholinguistic dimension, namely familiarity, transparency, ambiguity, knowledge, valence, and arousal, predict idiom judgment advantage, and do these effects vary across L1 and L2 speakers?

Support for RQ₂ would be shown by credible effects of specific deviation predictors, indicating that speaker-specific variation in each psycholinguistic dimension predicts idiom judgment advantage. Credible interactions with the language group would indicate that these dimension-specific effects differ between L1 and L2 speakers.

Methodology

Participants

Data were collected from 128 English speakers, including 71 L1 speakers (mean age = 22.6; 40 females) and 57 L2 speakers (mean age = 22.2; 38 females), all of whom were students at a U.S. university. L1 participants were U.S.-born native English speakers. The L2 participants were highly proficient L2 speakers with later immersion experience and met the following criteria: began acquiring English after age 3 and before age 10, at least three years of residence in English-speaking countries, and self-rated proficiency above 7 on a 9-point

scale. Most were exempt from standardized language tests or had advanced scores (IELTS > 7.5, TOEFL > 100). Their mean age of acquisition (AoA) was 5.6 years ($SD = 2.1$), with an average of 68.4 months ($SD = 43.5$) living in English-speaking environments. They reported using English for 59.2 hours per week ($SD = 21.2$) over an average of 63.1 months ($SD = 19.4$). Self-rated proficiency scores were: overall ($M = 7.8$, $SD = 0.8$), speaking ($M = 7.7$, $SD = 0.7$), listening ($M = 8.1$, $SD = 0.8$), reading ($M = 8.0$, $SD = 0.9$), and writing ($M = 7.7$, $SD = 0.9$). This profile differs from that of typical classroom-based L2 learners: participants were studying at a U.S. university, had substantial residence experience in English-speaking countries, and reported high proficiency. However, they should not be treated as simultaneous or balanced bilinguals, as all began acquiring English after age 3 and had, on average, over 16 years of English learning or use before full immersion. Participant demographics for the L2 group are summarized in Table 1.

[Insert Table 1 about here]

We conducted an exploratory L2-only participant-level analysis using two broad language-experience measures: length of residence in English-speaking countries and weekly English use. As reported in Appendix S1, neither measure reliably predicted participant-level mean idiom judgment advantage. This null finding should be interpreted cautiously because the analysis was exploratory and relied on broad self-reported measures rather than idiom-specific exposure.

Figure 1 shows that the L2 group was linguistically diverse, with Chinese-related backgrounds forming the largest subgroup. We therefore conducted an exploratory Chinese-English cross-language similarity analysis for this subgroup. As reported in Appendix S2,

Analysis 1, Chinese-English idiom overlap did not reliably predict judgment advantage, and the key ID effects remained qualitatively similar after including cross-language similarity. We also conducted a high-level exploratory comparison between the two largest L1-background subgroups, Chinese-related and Spanish backgrounds (Appendix S2, Analysis 2), which showed no evidence of subgroup differences in overall RT-based judgment advantage. [Insert Figure 1 about here]

Stimuli

A total of 4,356 English idioms were initially extracted from the *Oxford Dictionary of Idioms* (Siefring, 2005). Candidate expressions were retained only if they showed conventionalized figurative use. Purely literal expressions, non-conventional metaphors, and transparent comparison phrases were excluded. The stimulus set was controlled along length, affective, syntactic, semantic, and frequency dimensions. Idioms longer than five words were excluded to control for length. To reduce the possibility that idiom-level affective ratings were driven by single emotional words or by literal interpretations, two independent L1 English raters evaluated the valence and arousal of each idiom's component words on a 9-point scale. Idioms containing highly emotional words, defined as scores below 3 or above 7, were removed. The same raters then assessed the emotionality of each idiom's literal meaning, and idioms with extreme literal-meaning emotionality scores were excluded. Next, they rated the figurative meanings of the remaining idioms. Based on these ratings, idioms were assigned to nine groups formed by crossing valence (negative, neutral, positive) with arousal (low, medium, high) (See Table 2). This grouping was used to ensure that the final stimulus set covered a broad affective range, rather than being concentrated within a narrow

portion of the valence-arousal space. Then 25 or 26 idioms were randomly selected from each category, yielding 228 idioms for the experiment.

[Insert Table 2 about here]

For the 228 matched novel control phrases, the syntactic structure and syntactic category of each idiom were preserved by replacing one content word with an alternative from the same part-of-speech category and with comparable collocational properties (e.g., *kick the bottle* from *kick the bucket*). Replacement words were selected to be semantically similar to the original words where possible, and both the selected words and resulting control phrases were checked by the same L1 English raters for semantic plausibility and meaningfulness. Replacement words were also selected to be affectively similar to the original words where possible, and the same raters confirmed that both the selected words and resulting control phrases fell within the neutral valence/arousal range of 3-7 on the 9-point scales. Control phrases with COCA frequency values above 3 were excluded to minimize phrase-level frequency effects. In addition, 150 semantically meaningless filler phrases were constructed by randomly combining two to five words drawn from the idiom and control sets.

Procedure

The study comprised a phrase judgment task and a norming study, carried out in that order. In the phrase judgment task, 228 idioms, 228 matched novel phrases, and 228 fillers were divided into four lists (57 idioms, 57 controls and 57 fillers each). L1 and L2 participants were evenly assigned to one list. To avoid the co-occurrence of idioms and their controls, each pair was split into two sets (Set 1: 28 pairs; Set 2: 29 pairs). Each participant

completed two blocks: Block A contained idioms from Set 1 and controls from Set 2, and Block B reversed this pairing. Block order was counterbalanced. After informed consent and instructions, participants completed five practice trials with feedback, then proceeded to the main task without feedback. Each trial began with a 500-ms fixation, followed by a phrase. Participants judged its meaningfulness by pressing “F” (meaningful) or “J” (meaningless) within 4,000 ms. This task was designed to assess how efficiently participants recognized phrases as meaningful. Accordingly, RTs were interpreted as indexing phrase-level meaning recognition and decision efficiency, rather than a specific subprocess such as early idiom recognition, figurative meaning retrieval, or semantic integration. Accuracy and RTs were recorded using DMDX (Forster & Forster, 2003). The trial order was randomized, with a short break provided halfway.

The norming task followed the phrase judgment task. The norming task was designed to obtain speaker-specific ratings of psycholinguistic dimensions and to derive item-level average norms based on the mean ratings of L1 speakers. Each L1 and L2 participant rated the same idioms that they had responded to in the phrase judgment task on six psycholinguistic variables using a 9-point Likert scale (see Appendix S3 for more details). Valence and arousal were rated using the same dimensions and scale anchors as in stimulus selection, but by the experimental participants rather than the independent L1 raters. These follow as below:

(1) Familiarity (Bulkes & Tanner, 2017; Citron et al., 2016; Libben & Titone, 2008): How often the idiom had appeared to participants, rated from 1 = “I have never heard or read this idiom”, 5 = “I have sometimes heard or read this idiom”, to 9 = “I have very often heard

or read this idiom”.

(2) Transparency (Citron et al., 2016): The extent to which the idiom’s figurative meaning could be inferred from its parts, rated from 1 = “not at all”, 5 = “to some extent”, to 9 = “very easily”.

(3) Ambiguity (Bonin et al., 2013; Bulkes & Tanner, 2017; Libben & Titone, 2008): How likely the idiom could be interpreted literally across contexts, rated from 1 = “not at all”, 5 = “possibly”, to 9 = “almost always”.

(4) Knowledge (Bulkes & Tanner, 2017; Citron et al., 2016; Libben & Titone, 2008): How well participants knew the figurative meaning of idioms, rated from 1 = “not at all”, 5 = “approximately”, to 9 = “completely”. This measure captures participants’ subjective confidence in their idiom knowledge rather than an objective verification of meaning accuracy. Although this measure does not test objective definition accuracy, it captures speakers’ perceived access to idiom meaning, a factor that may shape how readily a phrase is recognized as meaningful and available during online judgment tasks.

(5) Valence (Citron et al., 2016; Stadthagen-Gonzalez et al., 2017; Warriner et al., 2013; Yao et al., 2017): How positive or negative the idiom felt, rated from 1 = “completely unhappy”, 5 = “neutral”, to 9 = “completely happy”.

(6) Arousal (Citron et al., 2016; Warriner et al., 2013): the extent to which it made participants feel calm or excited, rated from 1 = “very calm”, 5 = “moderately arousing”, to 9 = “very excited”.

Data Analysis

Table 3 summarizes the analytic workflow, with Steps 1-7 serving as preparatory

analyses and Steps 8-9 addressing RQ₁ and RQ₂.

[Insert Table 3 about here]

Descriptive Analysis

Data preprocessing

Norming and RT data yielded 7,012 observations (see Appendix S4 for raw data). We excluded participants whose ratings showed excessive consistency across dimensions, defined as assigning the same score to >80% of items, but no participants met this criterion. For averaged norming scores, outlier responses within each psycholinguistic dimension were removed using a ± 3 *SD* threshold around the mean. Outliers were rare, occurring only in arousal ($N = 23$) and knowledge ($N = 11$), and accounted for less than 1% of observations ($N = 3,819$ per dimension), indicating strong overall data integrity. For RT analyses, incorrect trials (Accuracy = 0) were excluded so that RTs reflected idiom comprehension. We also excluded two anticipatory outliers using predefined criteria: idiom or control RTs below 200 ms, or RTs more than 3 *SD* below a participant's mean RT, computed separately by trial type.

Rating consistency

To assess rating consistency across participants within each language group, we computed intra-class correlation coefficients (ICC_{2k} and ICC_{3k}) for L1 and L2 groups separately across all six psycholinguistic dimension ratings. Unlike Pearson correlations, which measure associations between variables or rating sets, ICCs estimate how consistently multiple raters rate the same items. Higher ICCs indicate greater agreement and more reliable item-level ratings. ICC_{2k} estimates reliability when raters are treated as sampled from a broader population, whereas ICC_{3k} estimates consistency among the specific raters in this

study.

L1 speakers showed high reliability across dimensions, with all ICCs exceeding .96 and the highest ICC_{3k} values for familiarity (.98), transparency (.97), and knowledge (.98). L2 speakers also showed high reliability across dimensions (ICCs > .93) despite greater proficiency variability, with familiarity and transparency again showing the strongest consistency (ICC_{3kS} = 0.90). These results indicate stable and internally coherent item ratings in both groups. To assess external validity, we compared our L1 ratings with published idiom norms. For the 39 identical idioms in Bulkes and Tanner (2017), correlations for ambiguity ($r = .74$), familiarity ($r = .53$), and transparency ($r = .74$) indicated moderate to substantial agreement (Cicchetti, 1994). For 98 semantically comparable but structurally different idioms (e.g., *up one's alley* vs. *in one's wheelhouse*), correlations remained stable for ambiguity ($r = .61$), familiarity ($r = .50$), and transparency ($r = .64$).

Overall, the ICC analyses indicate reliable item-level consistency within each language group while preserving meaningful speaker-specific variability.

Descriptive analysis

Table 4 summarizes L1 and L2 descriptive statistics for the six psycholinguistic dimensions and behavioral measures of idiom judgment advantage.

[Insert Table 4 about here]

Correlations across dimensions

To examine relationships among the six psycholinguistic dimensions within each group, we computed pairwise Pearson correlations for ambiguity, arousal, transparency, familiarity, knowledge, and valence separately for L1 and L2 speakers.

In the L1 group (Figure 2a), familiarity and knowledge were strongly correlated ($r = .72$). Transparency showed small-to-moderate positive correlations with familiarity ($r = .25$) and knowledge ($r = .31$). Ambiguity was weakly related to other dimensions, showing a small positive correlation with transparency ($r = .17$) and near-zero to small negative correlations with familiarity ($r = -.05$) and knowledge ($r = -.17$). Arousal showed weak negative correlations with familiarity ($r = -.23$), knowledge ($r = -.22$), and transparency ($r = -.14$), but near-zero correlations with ambiguity ($r = -.02$) and valence ($r = -.03$). Valence was only weakly related to the other dimensions ($-.10 < r_s < .00$).

In the L2 group (Figure 2b), correlations largely followed the L1 pattern. Transparency was positively correlated with familiarity ($r = .35$) and knowledge ($r = .23$), and more strongly correlated with ambiguity than in L1 ($r = .33$). Familiarity and knowledge remained strongly correlated, though slightly less so than in L1 ($r = .67$). Arousal showed small negative correlations with familiarity ($r = -.19$), knowledge ($r = -.14$), and transparency ($r = -.20$), whereas valence showed small positive correlations with ambiguity ($r = .14$), transparency ($r = .19$), and familiarity ($r = .16$), but was unrelated to knowledge ($r \approx 0$).

[Insert Figure 2 about here]

Semantic representations of idioms

To assess L1-L2 similarity in idiom representations, we constructed idiom-by-idiom dissimilarity matrices from the six rated dimensions (ambiguity, arousal, transparency, familiarity, knowledge, and valence) and applied Mantel and Procrustes analyses. The Mantel analysis tested cross-group correlations in overall pairwise dissimilarity structure, whereas the Procrustes analysis assessed how closely the multidimensional idiom spaces of L1 and L2

speakers aligned after rotation, translation, and scaling.

Heatmaps of the idiom distance matrices for L1 (Figure 3a) and L2 (Figure 3b) speakers showed highly similar clustering patterns, with nearly identical hierarchical clustering on both axes. These similarities suggest that L2 speakers encode idiom meanings in a semantic structure closely aligned with that of L1 speakers. In both panels, the diagonal reflects zero distance between identical idioms, whereas off-diagonal blocks indicate clusters of semantically related idioms.

[Insert Figure 3 about here]

The Mantel test revealed a strong correlation between the L1 and L2 dissimilarity matrices (Mantel $r = .986$, $p < .001$, 9,999 permutations), indicating highly similar idiom organization across L1 and L2 groups. Procrustes analysis likewise showed strong alignment between the L1 and L2 multidimensional rating spaces after rotation, translation, and scaling ($r = .990$, sum of squares = .019, $p < .001$).

To assess within-group organization, we applied agglomerative hierarchical clustering to the idiom rating matrices. The resulting cophenetic correlation coefficient ($C = .674$) indicated moderately strong clustering, suggesting coherent semantic structuring in both groups. Together, these results show that L1 and L2 speakers organize idioms along psycholinguistic dimensions in highly similar ways, consistent with previous findings (Morid & Sabourin, 2024). This cross-group alignment supports the structural reliability of the rating data and provides a foundation for subsequent individual-level modeling.

Dispersion of individual ratings relative to group norms

Following prior research, we used mean ratings from L1 speakers as item-level

average norms for subsequent analyses (Milburn et al., 2021; Siyanova-Chanturia et al., 2011; Soto-Sierra & Ferreira, 2024). To visualize dispersion around these norms, we plotted individual ratings against L1 group means for each dimension separately for L1 (Figure 4a) and L2 speakers (Figure 4b).

[Insert Figure 4 about here]

In L1 speakers, familiarity, knowledge, and transparency showed particularly tight alignment, suggesting robust shared representations, whereas ambiguity and arousal showed greater spread, indicating higher subjectivity. L2 ratings followed the same general linear trend but were more dispersed than L1 ratings, suggesting greater inter-individual variability, especially for emotional properties and ambiguity. In contrast, familiarity, knowledge, and transparency showed stronger clustering, suggesting more reliable acquisition of these properties by L2 learners. Overall, both groups showed internal coherence, but L2 learners adhered less closely to group norms, particularly for properties requiring emotional or inferential integration.

Group-level convergence across psycholinguistic dimensions

Finally, the radar plot of dimension-wise differences (L2-L1; Figure 5) showed no apparent group differences across the six dimensions, suggesting group-level representational convergence.

[Insert Figure 5 about here]

Bayesian multilevel model

Hierarchical Bayesian linear mixed-effects models, implemented in the R package *brms* (Bürkner, 2017), were used to analyze ID effects on judgment advantage. This approach

accounted for crossed random effects of participants and items, captured IDs in subjective ratings and behavioral measures, and provided full posterior distributions for all parameters. Analyses included only correct trials. RT advantage was defined as the standardized difference between idioms and their matched novel controls, allowing item-level subtraction to serve as the dependent measure and control for item-specific variability. This measure captures faster responses to idioms than to matched novel control phrases in the meaningfulness judgment task during real-time processing, rather than indexing a specific subprocess. To ensure that higher scores indicated greater judgment advantage, each item's value was derived by subtracting idiom RTs from their corresponding literal controls. All psycholinguistic predictors were z-transformed before analysis.

For each of the six rated dimensions, namely familiarity, knowledge, transparency, ambiguity, valence, and arousal, two predictors were computed: (1) the L1-based item-level normed rating, calculated as the mean rating for each item across L1 speakers. This predictor followed the common practice of using L1 ratings as relatively stable idiom norms (e.g., Libben & Titone, 2008) and provided a shared baseline for calculating speaker-specific deviations in both groups. (2) the participant-specific deviation from this norm (deviation scores), which indexed the extent to which a given speaker's individual perception or experience of that idiom differed from the item-level average norm. These predictors allowed us to separate item-level normed ratings from speaker-specific variation.

Two models tested ID effects on idiom judgment advantage. The Norm-Based Model included language group, the six L1-based group-mean psycholinguistic ratings, their interactions with language group, and random intercepts for participants and items. The Full

ID Model additionally included standardized participant-specific deviation scores for each dimension and their interactions with language group, testing whether individual-level ratings improved prediction and whether these effects differed between L1 and L2 speakers. Both models used weakly informative priors: Normal(0, 1) for fixed effects and Student- $t(3, 0, 0.5)$ for random effects and residuals. Four MCMC chains were run for 8,000 iterations, with convergence confirmed by $R\text{-hat} = 1.00$ and effective sample sizes $> 1,000$.

RQ₁ asked whether deviations in speaker-specific perceptions from item-level norms improved overall prediction of idiom judgment advantage beyond item-level norms, and whether this contribution differed between L1 speakers and highly proficient L2 speakers. To address RQ₁, we assessed model performance using leave-one-out cross-validation, examining differences in expected log predictive density (ELPD) and LOOIC, complemented by Bayesian R^2 to evaluate explained variance. RQ₂ asked how speaker-specific deviations in each psycholinguistic dimension predicted idiom judgment advantage and whether these effects differed between L1 speakers and highly proficient L2 speakers. To address RQ₂, we examined the posterior estimates of the deviation scores as well as their interaction terms with language group in the full model.

Results

As shown in Table 5, the deviation model outperformed the average-only model in the full sample ($\Delta\text{ELPD} = 1503.8$, $SE = 58.0$), with LOOIC decreasing from -1825.6 to -4833.2. Bayesian R^2 increased from 0.772 to 0.875 ($\Delta R^2 = 0.103$), indicating ~10% additional variance explained in RT advantage. Overall, modeling participant-specific deviations improved fit beyond item-level average norms.

[Insert Table 5 about here]

Separate analyses by language group showed the same pattern (Table 6). For L1 speakers, the deviation model markedly improved fit ($\Delta\text{ELPD} = 919.1$, $SE = 40.7$) and increased Bayesian R^2 from 0.773 to 0.883 ($\Delta R^2 = 0.110$). For L2 speakers, improvements were smaller but robust ($\Delta\text{ELPD} = 584.7$, $SE = 41.3$), with R^2 rising from 0.735 to 0.845 ($\Delta R^2 = 0.110$).

[Insert Table 6 about here]

As shown in Table 7 and Table 8, the Bayesian multilevel model indicated that L2 speakers showed smaller idiom judgment advantages than L1 speakers ($b = -0.268$, 95% CrI [-0.276, -0.260]).

For the effects of L1-based item-level norms, Familiarity positively predicted idiom judgment advantage in both groups (L1: $b = 0.199$, 95% CrI [0.129, 0.269]; L2: $b = 0.208$, 95% CrI [0.138, 0.279]). Valence and Arousal norms were also facilitatory in both groups (Valence: L1: $b = 0.088$, 95% CrI [0.063, 0.114]; L2: $b = 0.061$, 95% CrI [0.036, 0.087]; Arousal: L1: $b = 0.087$, 95% CrI [0.060, 0.115]; L2: $b = 0.025$, 95% CrI [0.003, 0.054]), but these effects were reliably weaker in L2 speakers (Valence: $b = -0.027$, 95% CrI [-0.036, -0.018]; Arousal: $b = -0.062$, 95% CrI [-0.071, -0.053]). Knowledge norms did not show a significant main effect in either group, but the interaction analysis indicated that their effect was reliably stronger in L1 than in L2 speakers ($b = -0.111$, 95% CrI [-0.134, -0.088]). In contrast, Transparency and Ambiguity norms were inhibitory in both groups (Transparency: L1: $b = -0.110$, 95% CrI [-0.143, -0.079]; L2: $b = -0.046$, 95% CrI [-0.078, -0.014]; Ambiguity: L1: $b = -0.164$, 95% CrI [-0.195, -0.134]; L2: $b = -0.149$, 95% CrI [-0.180, -

0.119)), with both inhibitory effects attenuated in L2 speakers (Transparency interaction: $b = 0.064$, 95% CrI [0.054, 0.074]; Ambiguity interaction: $b = 0.015$, 95% CrI [0.006, 0.025]).

Speaker-specific perception deviations from item-level average norms explained additional variance. Valence deviations significantly predicted idiom judgment advantage in both L1 speakers ($b = 0.006$, 95% CrI [0.000, 0.012]) and L2 speakers ($b = 0.051$, 95% CrI [0.041, 0.062]), with a reliably stronger effect in L2 speakers ($b = 0.046$, 95% CrI [0.034, 0.057]). Likewise, positive Transparency deviations facilitated idiom judgment advantage only in L2 speakers ($b = 0.065$, 95% CrI [0.054, 0.075]) and were significantly stronger than in L1 speakers ($b = 0.069$, 95% CrI [0.057, 0.081]). Ambiguity deviations, by contrast, showed a positive effect in both L1 ($b = 0.128$, 95% CrI [0.122, 0.134]) and L2 speakers ($b = 0.020$, 95% CrI [0.012, 0.028]), with the effect significantly larger in L1 than in L2 speakers ($b = -0.108$, 95% CrI [-0.118, -0.099]).

[Insert Table 7 about here]

[Insert Table 8 about here]

Discussion

General ID effects

The idiom judgment advantage indexes how efficiently participants recognized idioms as meaningful relative to matched novel controls, rather than a specific subprocess such as early recognition, figurative-meaning retrieval, or semantic integration. IDs in idiom-related psycholinguistic dimensions, captured through speakers' own ratings, significantly shaped judgment advantages in both L1 and L2 groups. This suggests that although these dimensions appear stable at the group level, aggregate norms do not fully capture experience-

based individual variability. Speaker-specific ratings are therefore more sensitive to item- and speaker-specific factors and prevent subtle experiential differences in phrase meaningfulness judgments from being obscured. These findings echo prior L1 and L2 research showing that IDs shape language processing (e.g., Andrews & Hersch, 2010; Cacciari et al., 2018; Milburn et al., 2021; Tilmatine et al., 2021) and extend this work by showing that such variability also arises in idiom-related psycholinguistic dimensions. The results suggest that phrase-level meaning access is experience-sensitive. For L1 speakers, cumulative and culturally embedded idiom experience may support efficient meaning judgments by strengthening structural, semantic, and emotional associations (Bybee, 2010; Ellis, 2002). For highly proficient L2 speakers, judgment advantages also changed when speaker-specific ratings were incorporated, suggesting that these ratings capture meaningful psycholinguistic cues used as idiom representations that continue to develop.

The overall benefits of modeling IDs were greater in L1 than in L2 speakers. One possibility is that L1 ratings are more stable and differentiated, whereas L2 ratings often show flatter response profiles and weaker itemwise differentiation (Morid & Sabourin, 2024; Wolter & Gyllstad, 2011). Under this interpretation, L1 between-participant variability may more systematically reflect differences in the strength and accessibility of idiom representations, which may increase the explanatory value of speaker-specific deviations for judgment advantage. In highly proficient L2 speakers, subjective ratings may also reflect variability beyond representational strength, including strategic reliance on constituent-based reasoning, because L2 idiom experience can remain uneven despite high proficiency and immersion (Morid & Sabourin, 2024). This pattern is consistent with dual-route accounts, in

which L2 idiom processing may rely more on compositional analysis unless holistic access is strongly supported (Libben & Titone, 2008). It also helps extend representation-based accounts (Abel, 2003) originally developed with non-immersed L2 learners to highly proficient immersed L2 speakers, suggesting that L2 idiom representations may remain less robust or less richly specified under some conditions. More broadly, it is compatible with the Shallow Structure Hypothesis, which posits less automatically deployed structural representations in adult L2 processing (Clahsen & Felser, 2006). It is noted that the smaller overall predictive gain in L2 speakers than in L1 speakers does not imply uniformly weaker ID effects for each specific dimension, which should be examined separately.

IDs in speakers' perceptions of psycholinguistic dimensions

Several ID effects were consistent across L1 and L2 speakers. Familiarity, knowledge, and arousal showed reliable L1-based item-level norm effects, but no additional contribution from deviations in participant-specific perceptions from the item-level mean in either L1 or L2 speakers. This suggests that these norms already capture the core experiential and emotional properties driving idiom processing, providing robust estimates of lexical-semantic properties by smoothing idiosyncratic, transient, and task-related variability. In contrast, positive deviations from item-level average norms in ambiguity and valence predicted judgment advantages for both L1 and L2 speakers, indicating that individual variation in semantic uncertainty and affective evaluation carries explanatory value beyond group-level norms.

IDs in psycholinguistic dimensions also showed asymmetric effects across groups. Perceived valence exhibited stronger ID effects in L2 than in L1 speakers. This pattern is

consistent with greater reliance on analytic strategies and more variable emotional grounding in L2 speakers (Milburn et al., 2026; Pavlenko, 2012; Sheikh & Titone, 2016). IDs in Ambiguity, by contrast, showed a stronger positive effect in L1 than in L2 speakers, plausibly because more entrenched idiom representations in L1 speakers allow them to exploit ambiguity-related interpretive flexibility more efficiently during online comprehension.

Convergent effects of IDs across L1 and L2 speakers

Item-level average familiarity norms consistently facilitated idiom judgment advantages in both L1 and L2 speakers, consistent with the findings that higher familiarity supports faster access to phrase-level meaning (Cacciari & Tabossi, 1988; Conklin & Schmitt, 2012; Siyanova-Chanturia et al., 2011). Familiar idioms are more likely to have readily available phrase-level representations, which may facilitate meaningfulness judgments for both groups (Cacciari & Tabossi, 1988; Libben & Titone, 2008). However, incorporating speaker-specific perceptions of familiarity did not improve model performance in either L1 or L2 speakers. This suggests that item-level average familiarity norms already capture the cumulative effects of conventionalized usage and long-term exposure, particularly for culturally entrenched expressions. By contrast, speaker-specific familiarity ratings may reflect temporary biases and may lack the resolution needed to robustly track subtle item-level variation (Morid & Sabourin, 2024). Therefore, item-level average familiarity norms appear to function as stable population-level predictors, with limited added value from speaker-specific familiarity deviations.

Item-level average knowledge norms did not show a significant main effect in either group, but there was still a reliable group difference. The interaction indicated that item-level

knowledge norms had a weaker positive effect in L2 than L1 speakers, plausibly because L2 idiom self-rated subjective knowledge is less entrenched and semantically integrated, limiting rapid access to phrase-level meaning. This pattern suggests that even highly proficient L2 speakers with later immersion experience may show reduced semantic integration of idiom representations relative to L1 speakers, extending the Representation Deficit account (Abel, 2003) to highly proficient immersed L2 speakers and aligning with proposals that L2 comprehension relies on comparatively shallow or less automatically deployed structural representations under online constraints (Clahsen & Felser, 2006). Notably, adding self-rated speaker-specific knowledge ratings did not improve model fit for either group, suggesting that self-rated knowledge, unlike objective meaning accuracy, may be too coarse to capture the phrase-level meaning availability that supports efficient idiom judgments.

Item-level average valence norms exerted a significant facilitatory effect on idiom meaningfulness judgments in both L1 and L2 speakers, consistent with evidence that negative information delays attentional engagement relative to positive content, thereby slowing access to the figurative meaning of idioms (Kuperman et al., 2014). Yet the effect was attenuated in L2 speakers, likely due to reduced emotional resonance and encoding in an L2 (Pavlenko, 2012). However, incorporating speaker-specific deviations in valence ratings significantly improved model fit. In both groups, idioms judged as more emotionally positive than the item-level average valence norm showed more efficient phrase meaning activation. This suggests that speaker-specific valence ratings capture affective engagement that item-level average valence norms cannot represent. The effect appears in both L1 and L2 speakers because emotional evaluation is inherently personal: language speakers rely on their own

emotional interpretations during online phrase meaningfulness judgments, making individual variation informative even when overall emotional resonance differs.

Item-level average arousal norms positively predicted judgment advantages for idioms in both L1 and L2 speakers, suggesting that emotionally activating information may support task-relevant availability of phrase-level meaning (Citron, 2013). The effect was smaller for L2 than for L1 speakers, consistent with attenuated emotional resonance in an additional language (Pavlenko, 2012). Reduced emotional grounding in highly proficient L2 speakers with later immersion experience may reflect differences in the contexts and affective richness of language experience, as well as less embodied experience and weaker autobiographical associations than in the L1 speakers, which together can dampen arousal effects (Sheikh & Titone, 2016). This may occur because many L2 idioms were initially learned through less naturalistic input before full immersion, rather than through early-life, emotionally rich, and autobiographically grounded interactions. Notably, speaker-specific deviations in arousal did not predict idiom judgment advantage in either group. This pattern may indicate that arousal is largely stimulus-driven, with limited reliable between-participant variability, and that decontextualized rating tasks make arousal difficult to assess precisely.

Item-level average ambiguity norms significantly inhibited idiom judgment advantages in both L1 and L2 speakers, consistent with findings that ambiguity activates competing interpretations and reduces the efficiency of phrase meaningfulness judgments (Titone & Connine, 1999). More ambiguous idioms often lack a dominant, context-independent reading that can be readily retrieved during phrase-meaning judgments. When the dominant figurative meaning is less readily accessible, speakers may be more likely to

rely on literal or constituent-based processing, especially for idioms that are highly ambiguous or less familiar. This greater reliance on analytic processing may slow phrase-meaning decisions. The inhibitory effect was stronger in L1 speakers, likely because they may rely more readily on conventionalized figurative interpretations for familiar idioms (Cacciari & Tabossi, 1988; Libben & Titone, 2008; Titone & Connine, 1999; Swinney & Cutler, 1979) and are therefore more disrupted when semantic indeterminacy weakens access to a dominant figurative meaning. In contrast, L2 learners may be less likely to assume idiomaticity unless strongly cued (Siyanova-Chanturia et al., 2011). Many idioms may already be processed more like compositional phrases, making ambiguity less disruptive for L2 speakers' judgment routines. Importantly, IDs in speaker-specific ambiguity deviations positively predicted judgment advantages in both L1 and L2 groups, which was the opposite direction of the item-level average norm effect. One plausible explanation is that, at the individual level, higher ambiguity ratings index semantic richness rather than unresolved competition. Speakers who perceive an idiom as having more literal or alternative readings may have richer, more flexible representations that support phrase-level meaningfulness judgments. In the present task, this representational richness may facilitate judgment efficiency rather than hinder it. Item-level average ambiguity norms collapse speakers with genuinely unclear representations and those with richly elaborated ones. Thus, speaker-specific ambiguity deviations index disambiguation experience that item-level average norms cannot capture.

Item-level average transparency norms exhibited a general inhibitory effect on idiom judgment advantage in both L1 and highly proficient L2 groups, consistent with the dual-

route model (Libben & Titone, 2008), which posits that more transparent idioms may encourage constituent-based analysis and reduce reliance on whole-phrase cues during meaningfulness judgments. This item-level average transparency norm effect was more pronounced in L1 speakers than in L2 speakers, likely because constituent-based information interferes with their deeply entrenched idiomatic representations during the judgment task. In contrast, L2 speakers may use item-level average transparency as a compensatory mechanism when idiomatic knowledge is less entrenched. This also aligns with the Shallow Structure Hypothesis, which suggests that L2 speakers may rely more on surface-level or semantically driven cues, such as transparency, when fully entrenched linguistic representations are less automatically available (Clahsen & Felser, 2006).

Divergent ID effects across L1 and L2 speakers

The positive effect of speaker-specific ambiguity deviations tended to be stronger in L1 speakers than in L2 speakers. This pattern likely reflects group differences in representational strength and cue utilization. For L1 speakers, IDs in perceived ambiguity may be informative when idiom representations are well established, because alternative readings can be evaluated against a readily available figurative interpretation (Titone & Connine, 1999; Libben & Titone, 2008). Ambiguity may therefore serve as a useful cue rather than only a source of difficulty. For L2 speakers, the same facilitative pattern was present but weaker, perhaps because ambiguity ratings reflect more variable knowledge of how literal and figurative meanings are distributed across contexts (Milburn et al., 2021). Thus, perceived ambiguity may map less directly onto meaningfulness judgments in L2 speakers, where figurative access is more contingent on proficiency and the strength of instructed form-

meaning mappings (Libben & Titone, 2008; Clahsen & Felser, 2006).

In addition, speaker-specific valence deviations exerted a stronger positive impact on idiom judgment advantages in L2 speakers than in L1 speakers. This indicates that L2 speakers' speaker-specific perceptions of emotional content, though more variable, capture systematic differences in how emotion is engaged during phrase-level meaningfulness judgments. These divergences are not random fluctuations but reflect speaker-specific perceptions of emotional information. One explanation is that reduced emotional grounding in L2 learners, even highly proficient ones, may lead them to depend more on deliberate reasoning or contextual cues than on automatic emotional responses (Morid & Sabourin, 2024), making their speaker-specific valence ratings more predictive. This also suggests that ID effects in speaker-specific valence perceptions may be negligible in L1 speakers, where emotional representations are more entrenched and largely convergent, but may continue to shape variability in L2 speakers despite their advanced proficiency and immersion. This pattern aligns with experience-dependent accounts in which emotional processing is shaped by usage exposure and contextual diversity (Sheikh & Titone, 2016).

Importantly, speaker-specific transparency deviations positively predicted judgment advantage only in L2 speakers, with a stronger effect than in L1 speakers and in a direction opposite to the item-level average transparency norm effect. Individual L2 speakers showed a greater judgment advantage for idioms they rated as more transparent than the item-level norm, effectively overriding the baseline inhibitory effect of item-level transparency. This pattern indicates that speaker-specific perceptions of the association between an idiom's component words and its figurative meaning capture adaptive phrase judgment strategies in

L2 speakers. In L2 speakers, speaker-specific transparency ratings may reflect not only subjective perception but also the extent to which learners strategically draw on structural information based on their individual experience during phrase meaningfulness judgments. Viewing an idiom as more transparent may give L2 learners a more reliable basis for structure-driven parsing, resulting in stronger and directionally distinct speaker-specific perception effects. These findings also show that speaker-specific transparency effects are more pronounced in L2 speakers than in L1 speakers. For L1 speakers, holistic and automatized representations may dominate, leaving little room for additional speaker-specific transparency cues to influence meaningfulness judgments beyond the item-level average norm. In contrast, L2 learners rely more on analytic, composition-based processing when idiomatic representations are weak or still developing. This interpretation is consistent with recent visual-world evidence showing that even advanced L2 speakers may rely more strongly than L1 speakers on compositional processing during idiom processing (Milburn et al., 2026). Consequently, IDs in speaker-specific perceptions of transparency provide a more sensitive index of strategy use in L2 than in L1 idiom judgment.

Limitation

This study has several limitations. First, rating-based measures do not show that intrinsic idiom properties vary across speakers. Rather, they show that speakers differ in their perceptions of the same idioms, and that these differences predict judgment advantage beyond item-level norms. Future work could combine speaker-specific ratings with independent linguistic coding to distinguish intrinsic properties from subjective perceptions. Second, this study focused on transparency. Future work could examine compositionality as a

related but distinct dimension to clarify how idiom structure and perceived meaning inferability contribute to processing. Third, our L2 participants were highly proficient speakers with immersion experience in English-speaking environments. Future research should compare classroom L2 learners and early sequential bilinguals to examine language-use history. Fourth, our exploratory cross-language similarity analysis was limited to Chinese-background L2 speakers, the largest L1 subgroup. Future work should use larger samples from specific L1 backgrounds. Fifth, although the phrase judgment task indexes idiom judgment advantage, it cannot isolate subprocesses within idiom processing. Faster idiom responses may reflect phrase-form recognition, figurative-representation access, reduced semantic-integration demands, or decision-stage facilitation. Future work using eye tracking, self-paced reading, or electrophysiology could separate these subprocesses. Finally, speaker-specific knowledge was measured through self-ratings rather than definition production or meaning selection. Similar ratings may not indicate the same figurative interpretation, especially for idioms with multiple meanings and strong contextual dependence. Future work can combine self-ratings with objective meaning-accuracy measures to distinguish perceived from verified knowledge.

Conclusion

Taken together, these results show that idiom judgment advantages in English are shaped by psycholinguistic dimensions shared across L1 and L2 speakers, while also showing systematic ID effects. Although L1-based item-level ratings capture broad semantic, emotional, and structural regularities, IDs in valence, transparency, and ambiguity reveal additional representational diversity. These effects also differed between L1 and highly

proficient L2 speakers, suggesting that idiom judgment advantage reflects the interplay between entrenched linguistic conventions and individualized experiences shaped by exposure, representations, and processing strategies. Overall, idiom processing reflects the joint influence of shared psycholinguistic dimensions and speaker-specific experience, with their relative weight differing across L1 and L2 speakers.

Data Availability Statement

The supplementary and supporting materials associated with this study are available on OSF at <https://osf.io/4x2ju>.

Supplementary materials statement

Appendices are available at <https://osf.io/4x2ju>

References

- Abel, B. (2003). English idioms in the first language and second language lexicon: A dual representation approach. *Second Language Research*, 19(4), 329–358.
<https://doi.org/10.1191/0267658303sr226oa>
- Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General*, 139(2), 299–318. <https://doi.org/10.1037/a0018366>
- Bonin, P., Méot, A., & Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. *Behavior Research Methods*, 45, 1259–1271.
<https://doi.org/10.3758/s13428-013-0331-4>
- Bulkes, N. Z., & Tanner, D. (2017). “Going to town”: Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods*, 49(2), 772–783.

<https://doi.org/10.3758/s13428-016-0747-8>

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan.

Journal of Statistical Software, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

<https://doi.org/10.1017/CBO9780511750526>

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and*

Language, 27(6), 668–683. [https://doi.org/10.1016/0749-596X\(88\)90014-9](https://doi.org/10.1016/0749-596X(88)90014-9)

Cacciari, C., Corradini, P., & Ferlazzo, F. (2018). Cognitive and personality components

underlying spoken idiom comprehension in context: An exploratory study. *Frontiers in*

Psychology, 9, Article 659. <https://doi.org/10.3389/fpsyg.2018.00659>

Carrol, G., & Conklin, K. (2014). Getting your wires crossed: Evidence for fast processing of

L1 idioms in an L2. *Bilingualism: Language and Cognition*, 17(4), 784–797.

<https://doi.org/10.1017/S1366728913000795>

Carrol, G., & Conklin, K. (2017). Cross-language lexical priming extends to formulaic units:

Evidence from eye-tracking suggests that this idea “has legs”. *Bilingualism: Language*

and Cognition, 20(2), 299–317. <https://doi.org/10.1017/S1366728915000103>

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and

standardized assessment instruments in psychology. *Psychological Assessment*, 6(4),

284–290. <https://doi.org/10.1037/1040-3590.6.4.284>

Citron, F. M. M. (2012). Neural correlates of written emotion word processing: A review of

recent electrophysiological and hemodynamic neuroimaging studies. *Brain and*

Language, 122(3), 211–226. <https://doi.org/10.1016/j.bandl.2011.12.007>

- Citron, F. M. M., Weekes, B. S., & Ferstl, E. C. (2013). Effects of valence and arousal on written word recognition: Time course and ERP correlates. *Neuroscience Letters*, 533, 90–95. <https://doi.org/10.1016/j.neulet.2012.10.054>
- Citron, F. M. M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., & Jacobs, A. M. (2016). When emotions are expressed figuratively: Psycholinguistic and affective norms of 619 idioms for German (PANIG). *Behavior Research Methods*, 48(1), 91–111. <https://doi.org/10.3758/s13428-015-0581-4>
- Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, 27(1), 3–42. <https://doi.org/10.1017/S0142716406060024>
- Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29(1), 72–89. <https://doi.org/10.1093/applin/amm022>
- Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, 32, 45–61. <https://doi.org/10.1017/S0267190512000074>
- Cronk, B. C., & Schweigert, W. A. (1992). The comprehension of idioms: The effects of familiarity, literalness, and usage. *Applied Psycholinguistics*, 13(2), 131–146. <https://doi.org/10.1017/S0142716400005531>
- Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, 66(5), 843–863. <https://doi.org/10.1080/17470218.2012.720994>
- Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for

- theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2), 143–188. <https://doi.org/10.1017/S0272263102002024>
- Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1), 1–24. <https://doi.org/10.1093/applin/ami038>
- Ellis, R. (2004). Individual differences in second language learning. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 525-551). Blackwell Publishing.
- Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35, 116–124. <https://doi.org/10.3758/BF03195503>
- Gibbs, R. W., Jr. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.
- Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, 22(2), 154–169. <https://doi.org/10.1016/j.tics.2017.11.006>
- Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, 143(3), 1065–1081. <https://doi.org/10.1037/a0035669>
- Kyriacou, M., & Köder, F. (2024). Not batting an eye: Figurative meanings of L2 idioms do not interfere with literal uses. *Languages*, 9(1), Article 32. <https://doi.org/10.3390/languages9010032>
- Lewellen, M. J., Goldinger, S., Pisoni, D. B., & Greene, B. G. (1993). Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic

- categorization. *Journal of Experimental Psychology: General*, 122(3), 316-330.
<https://doi.org/10.1037/0096-3445.122.3.316>
- Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition*, 36, 1103-1121. <https://doi.org/10.3758/MC.36.6.1103>
- Milburn, E., & Warren, T. (2019). Idioms show effects of meaning relatedness and dominance similar to those seen for ambiguous words. *Psychonomic Bulletin & Review*, 26(2), 591–598. <https://doi.org/10.3758/s13423-019-01589-7>
- Milburn, E., Vulchanova, M., & Vulchanov, V. (2021). Collocational frequency and context effects on idiom processing in advanced L2 speakers. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 75(2), 169-180.
<https://doi.org/10.1037/cep0000214>
- Milburn, E., Vulchanova, M., Vulchanov, V., Saltzman, D., & Magnuson, J. S. (2026). Native speakers kick buckets, but learners kick doors: A comparison of native and nonnative idiom comprehension. *Memory & Cognition*, Advance online publication.
<https://doi.org/10.3758/s13421-025-01843-5>
- Morid, M., & Sabourin, L. (2023). Role of affective factors and concreteness on the processing of idioms. *Journal of Psycholinguistic Research*, 52(6), 2321–2338.
<https://doi.org/10.1007/s10936-023-10001-4>
- Morid, M., & Sabourin, L. (2024). Affective and sensory–motor norms for idioms by L1 and L2 English speakers. *Applied Psycholinguistics*, 45(1), 138–155.
<https://doi.org/10.1017/S0142716423000504>
- O'Reilly, D., Onysko, A., Rasse, C., Papitsch, L., Colston, H., & van der Horst, I. (2025).

- High ceilings and ingenuine allies: Tapping into the idiom meaning knowledge of first and second language speakers of English. *Language and Cognition*, 17, Article e72.
<https://doi.org/10.1017/langcog.2025.10029>
- Pavlenko, A. (2012). Affective processing in bilingual speakers: Disembodied cognition?. *International Journal of Psychology*, 47(6), 405–428.
<https://doi.org/10.1080/00207594.2012.743665>
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172. <https://doi.org/10.1037/0033-295X.110.1.145>
- Senaldi, M. S. G., & Titone, D. A. (2022). Less direct, more analytical: Eye-movement measures of L2 idiom reading. *Languages*, 7(2), Article 91.
<https://doi.org/10.3390/languages7020091>
- Sheikh, N. A., & Titone, D. (2016). The embodiment of emotional words in a second language: An eye-movement study. *Cognition and Emotion*, 30(3), 488–500.
<https://doi.org/10.1080/02699931.2015.1018144>
- Siefring, J. (Ed.) (2005). *The Oxford Dictionary of Idioms (2nd ed.)*. Oxford University Press.
- Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, 36(5), 549-569. <https://doi.org/10.1093/applin/amt054>
- Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27(2), 251-272. <https://doi.org/10.1177/0267658310382068>
- Soto-Sierra, V., & Ferreira, R. A. (2024). The influence of cross-language similarity and transparency on idiom knowledge in non-immersed L2 speakers. *System*, 122, Article

103287. <https://doi.org/10.1016/j.system.2024.103287>

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49(1), 111-123. <https://doi.org/10.3758/s13428-015-0700-2>

Steinel, M. P., Hulstijn, J. H., & Steinel, W. (2007). Second language idiom learning in a paired-associate paradigm: Effects of direction of learning, direction of testing, idiom imageability, and idiom transparency. *Studies in Second Language Acquisition*, 29(3), 449–484. <https://doi.org/10.1017/S02722263107070271>

Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, 18(5), 523–534. [https://doi.org/10.1016/S0022-5371\(79\)90284-6](https://doi.org/10.1016/S0022-5371(79)90284-6)

Tilmatine, M., Hubers, F., & Hintz, F. (2021). Exploring individual differences in recognizing idiomatic expressions in context. *Journal of Cognition*, 4(1), Article 37. <https://doi.org/10.5334/joc.183>

Titone, D. A., & Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbol*, 9(4), 247-270. https://doi.org/10.1207/s15327868ms0904_1

Titone, D. A., & Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, 31(12), 1655–1674. [https://doi.org/10.1016/S0378-2166\(99\)00008-9](https://doi.org/10.1016/S0378-2166(99)00008-9)

Titone, D. A., Columbus, G., Whitford, V., Mercier, J., & Libben, M. R. (2015). Contrasting bilingual and monolingual idiom processing. In R. R. Heredia & A. B. Cieślícka (Eds.),

Bilingual figurative language processing (pp. 171-207). Cambridge University Press.

<https://doi.org/10.1017/CBO9781139342100.011>

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4), 1191–1207.

<https://doi.org/10.3758/s13428-012-0314-x>

Whitford, V., & Titone, D. (2012). Second-language experience modulates first- and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic Bulletin & Review*, 19, 73–80.

<https://doi.org/10.3758/s13423-011-0179-5>

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, 32(4), 430–449.

<https://doi.org/10.1093/applin/amr011>

Yao, Z., Wu, J., Zhang, Y., & Wang, Z. (2017). Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, 49, 1374-1385. <https://doi.org/10.3758/s13428-016-0793-2>

Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations.

Studies in Second Language Acquisition, 40(4), 831–856.

<https://doi.org/10.1017/S0272263118000141>

Figure captions

Figure 1. The distribution of L1 backgrounds among the L2 participants.

Figure 2. Correlation matrix of psycholinguistic dimensions in (a) L1 speakers and (b) L2 speakers.

Figure 3. Idiom distance matrices for (a) L1 and (b) L2 speakers. The heatmaps show pairwise semantic distances among idioms in the six-dimensional rating space. Cooler colors indicate smaller distances, reflecting greater semantic similarity.

Figure 4. L1 (a) and L2 (b) participants' ratings for a given idiom were plotted against the item-level average norm for the same idiom across the six psycholinguistic dimensions.

Strong clustering around the identity line ($y = x$) suggests minimal individual deviation.

Figure 5. Radar plot illustrating the differences between advanced L2 speakers and L1 speakers across six psycholinguistic dimensions.