NOTION
Journal of Linguistics, Literature, and Culture

Volume 8, Number 2, November 2026
p. 380-395
http://journal2.uad.ac.id/index.php/notion/index

Cross-linguistic variation in Indonesian conceptual metaphor comprehension among BIPA learners: An exploratory study

Riki Nasrullah[1]*, Hishamudin Isam[2]
[1]Universitas Negeri Surabaya, Surabaya, Indonesia
[2]Universiti Utara Malaysia, Kedah, Malaysia

*Corresponding author: rikinasrullah@unesa.ac.id

Citation: Nasrullah, R., & Isam, H. (2026). Cross-linguistic variation in Indonesian conceptual metaphor comprehension among BIPA learners: An exploratory study. *NOTION: Journal of Linguistics, Literature, and Culture, 8*(2), 380-395. https://doi.org/10.12928/notion.v8i2.16747

ARTICLE INFO

Article History:
Received: 2026-06-15
Accepted: 2026-09-30

Keywords:
Conceptual metaphor theory
Cross-linguistic variation
Second language acquisition
BIPA
Figurative language comprehension

This work is licensed under a Creative Commons Attribution 4.0 International License

ABSTRACT

Conceptual metaphor comprehension in an additional language draws on mappings that may converge with or diverge from learners' first-language (L1) conceptual systems. This exploratory mixed-methods study examined L1-related variation and Indonesian proficiency in metaphor comprehension among 20 BIPA learners (n = 5 per L1 group: English, Mandarin, Japanese, and Arabic). Participants completed a 40-item multiple-choice metaphor task and think-aloud protocols. Participant-level accuracy differed across L1 groups, $F(3, 16) = 5.56$, $p < .01$, $\eta p^2 = .51$. Tukey HSD comparisons indicated higher accuracy for the English group than the Japanese and Arabic groups, whereas the remaining pairwise contrasts were not significant. Participant-level inspection suggested a possible tendency toward higher accuracy among C1 than B1 participants in the available sample, but proficiency was not tested inferentially because the L1-by-proficiency cells were small. The conceptual-equivalence classification used to construct the item set was independently validated by two native-speaker validators per L1, with high agreement across language-specific judgments (κ = .80–.86); however, the equivalent/non-equivalent distinction was not tested as a separate inferential factor. Think-aloud coding showed greater reliance on direct translation among Japanese- and Arabic-L1 participants and greater use of contextual inference among English-L1 participants; 30% of transcripts were double-coded with high agreement (κ = .84). Given the small sample, the findings are interpreted as preliminary between-L1 patterns rather than population-level or causal effects. The findings suggest potential value in contrastive, culturally contextualized metaphor instruction in BIPA and provide a basis for larger-scale research.

I. INTRODUCTION

Metaphor is a central mechanism of meaning-making because linguistic expressions can organize abstract experience through systematic conceptual mappings. Cognitive-linguistic accounts therefore treat metaphor as a structure of conceptual organization that is reflected in everyday language (Cienki, 2022; Gibbs, 2023; Lakoff & Johnson, 1980; Tong et al., 2021; Febriansyah, 2020; Fitria, 2022). Indonesian contains conventional figurative expressions such as *buah bibir* 'topic of conversation', *cuci tangan* 'avoiding responsibility', and *panjang tangan* 'stealing'. Their interpretation depends on lexical knowledge together with culturally conventionalized associations that connect a linguistic form to a figurative conceptual frame.

In Indonesian as an additional language, conceptual metaphor comprehension can involve cross-linguistic comparison between Indonesian expressions and conceptual resources available in the learner's L1. When an Indonesian metaphor has a familiar conceptual counterpart, prior L1 knowledge may facilitate access to its figurative interpretation. When the mapping is culturally specific or differs from patterns familiar in the L1, learners may depend more strongly on contextual inference, lexical analysis, or translation-based mediation (Boers & Muñoz-Basols, 2021; Carston & Yan, 2023; MacArthur & Alejo-González, 2024; O'Reilly & Marsden, 2021; Putri et al., 2025). Metaphor comprehension in an additional language consequently reflects an interaction among conceptual correspondence, linguistic familiarity, contextual information, and cross-linguistic transfer.

Research on second-language figurative competence has repeatedly linked metaphor comprehension to conceptual familiarity, cross-language correspondence, and broader language proficiency. Part-of-speech labels that indicate lexical relationships that are already widely known do not pose complex interpretive challenges for learners; however, when it comes to conceptual maps that are not yet familiar, students are likely to rely on independent word analysis or translation (Citron et al., 2020; Rizzato, 2021; Yu & Wang, 2019). General abilities can lead to metaphor-related performance, although competence also depends on exposure and experience with the target language (O'Reilly & Marsden, 2023). Research in this area remains limited, particularly regarding the Indonesian language, as the psycholinguistic literature on metaphor processing in a second language is largely written in English and other languages that have historically been widely used in research.

This gap is highly relevant to the field of BIPA (Indonesian for Foreign Speakers). In the Indonesian language, metaphorical meaning is inextricably linked to local cultural categories, social conventions, and prevailing patterns of conceptualization. Previous BIPA studies have examined curricula and proficiency frameworks, learner writing, instructional materials, cross-linguistic transfer, and teaching strategies (Fitria, 2023; Ningsih, 2023; Permadi et al., 2022; Pratiwi et al., 2023; Subandiyah et al., 2023). However, comparative evidence on how conceptual metaphors is understood across different first language (L1) backgrounds remain

very limited. This study therefore examines speakers whose L1s are English, Mandarin, Japanese, and Arabic. These four languages serve as the basis for an exploratory comparison. Each group of speakers brings a unique linguistic and cultural repertoire, allowing researchers to observe whether patterns of conceptual correspondence emerge across various L1 backgrounds, rather than being limited to a single language pair. No a priori ordering was assumed among the English-, Mandarin-, Japanese-, and Arabic-L1 groups.

Accordingly, this study addresses two research questions: (RQ1) How does L1 background relate to accuracy and interpretive strategies in Indonesian conceptual metaphor comprehension? (RQ2) What descriptive proficiency-related patterns are observed among B1- and C1-level participants? In line with this design, two exploratory expectations are taken as points of reference. First, participants whose L1 offers fewer conceptual correspondences with the Indonesian metaphors may find them more difficult to comprehend. Second, C1-level participants may display higher comprehension accuracy than B1-level participants. Accordingly, these expectations guide exploratory interpretation rather than confirmatory hypothesis-testing at the population level.

II. METHOD

This study employed an exploratory, mixed-methods comparative design. Two primary instruments are used: a controlled comprehension task and verbal protocol data. Quantitative analysis focuses on accuracy levels in metaphor comprehension. Meanwhile, qualitative analysis examines the interpretive strategies verbalized by participants during the think-aloud task. The aim of this design is not to measure causal effects, but rather to identify preliminary cross-linguistic patterns that can be further tested with larger samples.

Research Design

In the quantitative component, participant-level responses were compared across four first-language (L1) groups. The 40-item instrument was constructed from two categories of validated items: expressions with identifiable conceptual equivalents in the participants' L1s, and expressions lacking direct conceptual equivalents. These two categories served as the basis for stimulus selection and descriptive interpretation of cross-linguistic patterns, rather than as a separately tested inferential factor. Accordingly, the inferential analysis did not estimate conceptual equivalence as a separate within-participant factor. Accuracy served as the primary metric for cross-L1 comparison. The qualitative component employed a think-aloud protocol to examine translation, contextual inference, literal interpretation, and the use of L1 conceptual resources. Given the small group sizes, the study is exploratory in nature and was not designed to establish causal relationships.

Participants

The study comprised 20 participants of the BIPA program who were equally assigned to four groups. Each group had members representing one native language, which were English, Mandarin, Japanese, and Arabic, respectively. Therefore, every participant's mother tongue (L1) was only covered by n = 5. We selected the four native languages (L1s) based on purpose to represent maximum language and culture diversity, so they can be well compared with Indonesian as a kind of exploring language. All participants were at B1 (intermediate) or C1 (advanced) level, which were the outcome of a university-level placement test of BIPA. However, because the combination of L1 and proficiency level was represented by a limited number of participants, comparisons between levels are discussed purely descriptively, rather than as a confirmatory analysis of cross-level effects.

The selection process is based on predetermined criteria. Participants must have at least one year of continuous experience with the Indonesian language. In addition, they must not have grown up in a bilingual environment, nor have they ever lived in Indonesia for more than two years. These requirements were necessary to minimize as much as possible the extreme variations in bilingual experience and long-term cultural immersion. Nevertheless, the diversity among BIPA participants is maintained.

The participants in this study were purposively recruited from the BIPA program at the Faculty of Languages and Arts, Surabaya State University. While participants who fit the criteria were identified, the decision also involved considering the subjects' availability. More importantly, participant recruitment was planned to ensure representation of the four targeted first language (L1) backgrounds. As a result, the final sample included 20 subjects who were equally divided into four L1 groups: English, Mandarin, Japanese, and Arabic, with five participants per group. The subjects' ages ranged from 21 to 37 years (M = 27.4; SD = 4.6), of whom 11 were males and 9 females. The participants were from different educational levels, ranging from undergraduates and bachelor's degree holders to graduate students from various disciplines.

To determine the level of learners, an internal BIPA placement test is one of the ways this language course makes use of. Besides evaluating receptive and productive skills, reading comprehension, listening comprehension, vocabulary, grammar, and written language use also come into play when the test is carried out. CEFR is an international language proficiency scale upon which a person's level of language proficiency can be assessed. Hence, the outcomes of the language program's internal placement test should reflect learners' actual levels when compared to each other using CEFR. Based on the results of the institutional test and the applicable proficiency descriptors, these skills are then grouped into two levels: intermediate (B1) and advanced (C1). The B1 category represented learners who could understand the main points of familiar discourse and communicate effectively in common

academic and social situations. In contrast, the C1 category represented learners who could process more complex texts and use Indonesian flexibly and effectively across academic and social contexts. Given the small number of participants within the L1-by-proficiency cells, proficiency-related comparisons were treated as exploratory and descriptive rather than as confirmatory inferential effects.

Materials and instruments

The metaphor comprehension instrument contained 40 Indonesian expressions. Twenty items were categorized as having a conceptual equivalent in the participant's L1, and 20 were categorized as lacking a direct conceptual equivalent. Conceptual equivalence was operationalized as correspondence between the figurative conceptual relation expressed in Indonesian and a conventionalized metaphorical relation available in the relevant L1. The classification was established through the multilingual validation procedure reported below.

For example, Indonesian "*cuci tangan*" literally means 'wash hands' and conventionally refers to avoiding responsibility. English provides a related mapping in wash one's hands of something, whereas an L1 that does not conventionalize a comparable association may require a different inferential route. Item selection therefore considered conceptual correspondence, cross-linguistic interpretability, and the possibility of L1-based transfer.

The classification of the 40 Indonesian metaphor items according to the presence or absence of conceptual equivalents in participants' L1s was subjected to expert and native-speaker validation before the main data collection. Eight validators were involved, with two validators representing each of the four target language backgrounds: English, Mandarin, Japanese, and Arabic. All validators were native speakers of the language they evaluated and had academic or professional experience in linguistics, language education, translation, or cross-cultural communication. Several had prior experience working with Indonesian as an additional or foreign language.

Each validator independently examined the Indonesian metaphorical expression, its intended figurative meaning, and its possible counterpart in the relevant L1. An item was classified as having a conceptual equivalent when the corresponding L1 contained a conventional or readily recognizable metaphorical expression based on a comparable source–target conceptual mapping. An item was classified as having no direct conceptual equivalent when the figurative meaning could not be accessed through a comparable conventionalized mapping in the L1 and required a substantially different lexical or conceptual representation. The validators completed the classification independently to minimize consensus bias during the initial judgment stage.

Inter-rater reliability was calculated separately for each first language (L1), with two raters evaluating the test items independently. The results were analyzed using Cohen's Kappa statistic. There was a fairly high level of agreement across all four language groups: $\kappa = 0.86$ in

English, κ = 0.82 in Mandarin, κ = 0.84 in Japanese, and κ = 0.80 in Arabic. In cases where the two raters did not agree, they and the research team members discussed the problematic items. These items were assigned a category only after their conceptual equivalence had been confirmed. On the other hand, items whose conceptual equivalence remained in doubt were revised or excluded from the final stimulus set. The result of this process was the creation of a stimulus set containing 40 metaphor items, of which 20 were assigned a conceptual equivalence category and the other 20 had no direct equivalent in the first language (L1).

There are two main instruments used. The first instrument is a metaphor comprehension task consisting of 40 multiple-choice questions, in which participants are asked to select the most appropriate meaning of a metaphor from four options. The second instrument is the "think-aloud" method, in which participants are required to verbalize their thoughts while answering the metaphor questions. Metaphor comprehension serves as a quantitative criterion for determining differences between groups.

Each session lasted between 60 and 75 minutes and took place under controlled conditions. Short breaks were given to participants between tasks to reduce fatigue. The study employed two complementary methods: measuring comprehension and administering a verbal protocol. These two sources of information enabled behavioral patterns to be compared with participants' interpretations, as expressed through the verbal protocol. However, the study's ultimate goal remained exploratory.

This study has been approved by the Research Ethics Committee of Surabaya State University, with registration number 0023/UN38.III.1/DL.01.02/2025. First, all participants were given a thorough explanation of the procedures they would undergo. They were then informed that participation in the study was voluntary and that they could withdraw at any time without facing any negative consequences. After this briefing, all participants were required to sign a consent form. Each participant was assigned a unique identification code. We guarantee that the data will be kept secure and used solely for research purposes.

Data analysis

The units of quantitative analysis are individuals, each of whom provides responses to 40 metaphor items. These responses are then combined to give each individual an accuracy score, which is used to ensure that responses provided by the same individual cannot be considered independent prior to L1-L1 inference. Consequently, the model tests overall accuracy rather than the effects of individual conceptual equivalence categories.

Descriptive statistics illustrate the extent to which metaphor comprehension is accurate based on L1 group membership. Proficiency levels are descriptive rather than inferential. To conduct exploratory pairwise comparisons across groups, a one-way ANOVA was performed on the accuracy scores of participants across the four L1 groups (N = 20; n = 5 per group). Partial eta-squared was then used as a measure of effect size from the comprehensive test.

Following the identification of a statistically significant difference in the comprehensive test, Tukey's HSD post hoc contrasts were employed to conduct exploratory comparisons between groups. No inferential tests were conducted for the ability variable due to the small sample size of the L1 groups based on ability. As the results of the formal Shapiro–Wilk and Levene's tests were not included in the archived analysis records, the results of the ANOVA and post hoc contrasts can only be considered exploratory and should be interpreted alongside effect sizes, descriptive statistics and qualitative observations. Retrospective estimation of internal consistency using Cronbach's alpha was not possible due to missing response matrices in the records.

A thematic analysis process was conducted on the "think-aloud" protocol, involving transcription, initial coding, category formation and interpretation of iterative processing strategies. Data coding was conducted using four procedures: direct translation, contextual inference, literal interpretation and conceptual equivalents from the first language (L1). The percentage of processing strategies was determined based on the total number of L1 strategies coded in the selected protocol segments rather than the number of participants since some participants provided more than one coded strategy.

Two researchers who are experts in applied linguistics and BIPA research independently coded the 'think-aloud' protocol. Approximately 30 per cent of the transcripts were double-coded to assess consistency in coding. The level of agreement between the coders was high, with a Cohen's Kappa coefficient of 0.84 being found. Any disagreements were resolved through discussion, after which the agreed coding scheme was applied to the remaining transcripts. The participants' identities were concealed through the use of participant codes. Therefore, no identifying information is included in the excerpts below.

III. RESULTS AND DISCUSSION

Accuracy in Indonesian metaphor comprehension

The accuracy with which metaphors were understood varied among the four groups of native speakers (L1). Native English speakers achieved the highest accuracy rate, with an average score of 77.5%. Next were Mandarin speakers, with an average score of 67.5%, followed by Japanese speakers, with an average score of 65%, and finally Arabic speakers, with an average score of 62.5%. It is important to note that there were only five participants in each group. This small sample size does not provide accurate data on all groups of native speakers.

A one-way ANOVA on participant-level accuracy scores indicated between-group variation, $F(3, 16) = 5.56$, $p = .008$, $\eta p^2 = .51$. In the Tukey HSD post-hoc test, the English group's results were significantly different from those of the Japanese group, with an adjusted p-value of 0.027. The English group's results were significantly different from those of the Arabic group as well, with an adjusted p-value = 0.008. The comparison of English and Mandarin groups, however, was not statistically significant. The adjusted p-value for these two groups was 0.092.

Other pairwise contrasts also showed no significant differences, with all adjusted p-values ≥ 0.595. Given n = 5 per group and the absence of retained formal assumption-test outputs, these inferential results are treated strictly as exploratory and should be replicated with larger samples.

Table 1. Mean accuracy and confidence intervals of Indonesian metaphor comprehension by L1 background

L1	Mean Accuracy (%)	Standard Deviation	Sample Size (n)	95% CI Lower Bound	95% CI Upper Bound
English	77.5	5.2	5	71.04	83.96
Mandarin	67.5	6.1	5	59.93	75.07
Japanese	65.0	6.5	5	56.93	73.07
Arabic	62.5	7.0	5	53.81	71.19

Table 1 summarizes the participant-level means and dispersion. The English group had the highest mean and the smallest standard deviation (SD = 5.2), whereas the Arabic group had the lowest mean and the largest standard deviation (SD = 7.0). The overlapping confidence intervals for several groups reinforce the need to distinguish the significant omnibus pattern and selected pairwise contrasts from broader claims about all four L1 backgrounds.

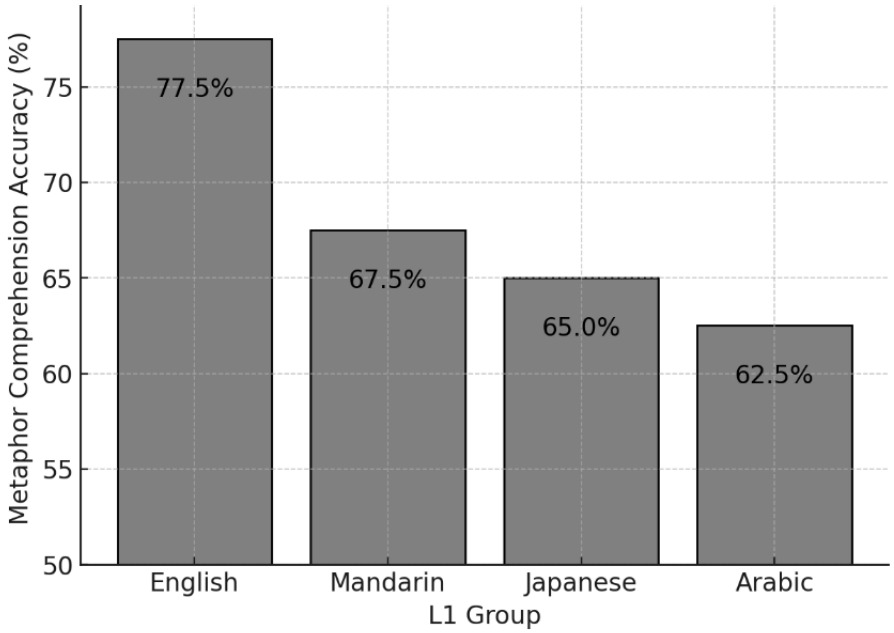


Figure 1. Metaphor Comprehension Accuracy Across L1 Groups of BIPA Learners

Figure 1 visualizes the descriptive ordering of L1-group means. The inferential analysis aggregated responses across all 40 items and therefore does not estimate a separate effect of conceptual-equivalence category. Accordingly, the validated equivalent/non-equivalent distinction is used as an item-construction and interpretive framework rather than as evidence

of a directly tested within-participant effect. The small sample and unmeasured exposure-related differences further preclude causal interpretation of the group ordering.

Proficiency-related patterns in metaphor comprehension

The proficiency comparison was descriptive because the L1-by-proficiency cells were too small for stable cell-level estimation or factorial inference. Inspection of participant-level scores suggested a possible tendency for C1 participants to achieve higher metaphor-comprehension accuracy than B1 participants in the available sample. Given the very small size of the subgroups, the mean and standard deviation for each cell is not included in the report. The apparent proficiency patterns should be interpreted only as preliminary findings, not as evidence of a standalone proficiency effect.

These proficiency-related trends must also be interpreted within the context of the participants' learning histories. In an institutional setting, higher proficiency tends to coincide with longer or more intensive exposure to Indonesian, superior vocabulary mastery, and broader experience with Indonesian cultural conventions. However, these dimensions were not modeled separately in the current study's dataset.

Overall, the patterns observed in the participants' responses indicate that proficiency in Indonesian can lead to better figurative language use. However, the design of the experiment did not permit distinguishing between proficiency levels and the duration, intensity or quality of exposure to the language. A larger study that includes groups with similar proficiency levels and measures language exposure is needed to estimate the contribution of each of these factors. The patterns associated with proficiency levels complement the analysis relating to first language (L1). However, these patterns do not suggest an interaction between the two.

Think-aloud analysis: Cognitive strategies in metaphor processing

Analysis of the 'think-aloud' data revealed three strategies that were used repeatedly in the encoded protocol: direct translation, contextual reasoning and literal interpretation. It is important to note that the percentages provided below indicate the frequency with which these strategies were used by members of each first language group, rather than the percentage of participants who used them. The reason for this is clear: a participant may use a single strategy repeatedly when analyzing the verbal protocol. Therefore, these results reflect the proportional distribution of these strategies among members of each L1 group.

Direct translation accounted for 60% of coded strategy instances in the Arabic group and 50% in the Japanese group. The corresponding proportions were 45% for the Mandarin group and 30% for the English group. This pattern indicates that direct translation was more frequently represented in the coded protocols of some L1 groups. These percentages should therefore be interpreted as distributions of coded strategy instances rather than as participant-level prevalence.

Contextual inference showed the reverse descriptive pattern. Contextual inference accounted for 55% of coded strategy instances in the English group, 40% in the Mandarin group, 35% in the Japanese group, and 25% in the Arabic group. Literal interpretation accounted for 15% of coded strategy instances in each L1 group. In addition, it should be noted that this picture reflects the strategies employed in the coded protocols only and should rather be treated as the record of strategy application during the session of a group. So, it in no way can be taken to mean that L1 background controls the choice of strategies.

Table 2. Distribution of coded think-aloud strategy instances by L1 group

L1	Direct translation (%)	Contextual inference (%)	Literal interpretation (%)
English	30 (n = 6)	55 (n = 11)	15 (n = 3)
Mandarin	45 (n = 9)	40 (n = 8)	15 (n = 3)
Japanese	50 (n = 10)	35 (n = 7)	15 (n = 3)
Arabic	60 (n = 12)	25 (n = 5)	15 (n = 3)

This table is intended to compare the proportional distribution of strategies. The percentages were calculated within each L1 group based on the coded strategy examples from the analyzed think-aloud segments. In other words, these percentages do not represent a percentage of the total number of participants and do not imply prevalence at the participant level based on the absolute number of these examples.

The previously coded patterns were further reinforced by the participants' verbal comments. P03, a C1-level English speaker, stated, "I understood it from the context; I didn't need to translate it first." That is a clear example of contextual inference. P17 was in a different situation. This B1-level Arabic speaker explained, "I translated the expression into Arabic first, but the literal meaning didn't help me." This statement indicates that he performed a direct translation. Meanwhile, P12 (Japanese, B1 level) reported, "At first, I consider the literal meaning, then I use the surrounding sentences to guess the figurative meaning." The shift from literal interpretation to contextual inference is clearly apparent in their statement.

Interpretation of cross-linguistic and proficiency patterns

In relation to RQ1, the combined quantitative and qualitative patterns indicate preliminary between-L1 variation in Indonesian metaphor comprehension and strategy use. English-L1 participants showed the highest mean accuracy, while Japanese- and Arabic-L1 participants relied more frequently on direct translation. The post-hoc analysis narrows this interpretation: statistically detectable contrasts were observed for English versus Japanese and English versus Arabic, whereas the other pairwise differences were not significant. Because the inferential model did not directly contrast equivalent and non-equivalent item categories, these findings support a selective between-L1 pattern within a validated conceptual-equivalence framework

rather than a direct estimate of conceptual-equivalence effects or a uniform hierarchy across all four L1s.

Participant-level inspection suggested a possible descriptive tendency for higher institutional proficiency to coincide with greater metaphor-comprehension accuracy. Because the L1-by-proficiency cells were very small, the study does not estimate cell-specific effects or a proficiency-by-L1 interaction. In addition, institutional proficiency cannot be isolated from accumulated exposure, instructional intensity, lexical familiarity, or familiarity with Indonesian cultural conventions. In relation to RQ2, the observed proficiency pattern remains preliminary, descriptive, and mechanism-neutral.

The think-aloud distributions provide a plausible strategic account of these behavioral patterns. Direct translation was more frequent among Japanese- and Arabic-L1 participants, whereas contextual inference was more frequent among English-L1 participants. Such patterns are consistent with the use of L1 resources as conceptual mediation when an immediately accessible Indonesian interpretation is unavailable. They should, however, be interpreted as coded strategy tendencies rather than direct measures of underlying cognitive mechanisms.

These findings support previous research. Research on second language (L2) has shown that metaphoric competence is related to cross-linguistic correspondence, conceptual familiarity, vocabulary knowledge, and general ability (Boers & Muñoz-Basols, 2021; Citron et al., 2020; O'Reilly & Marsden, 2021, 2023; Rizzato, 2021). A way of looking at the issue is the Conceptual Metaphor Theory, which explains why mapping a source to a familiar target is the kind of operation that supports the interpreter in different languages (Lakoff & Johnson, 1980). Nevertheless, it should not be forgotten that this research has been from the point of view of BIPA (Indonesian for Foreign Speakers).

The frequency of direct translation tended to be higher in the Japanese and Arabic language groups. This reflects a greater reliance on first-language (L1) mediated interpretation, particularly for the items used in this study. In contrast, the English group relied more heavily on contextual inference. The reasons for this could vary—ranging from familiarity with specific mappings, stronger item-level lexical access, and differences in exposure, to a combination of these factors. The study's design did not distinguish between these explanations; therefore, this remains a productive area for larger-scale controlled research.

This pattern aligns with previous research. Metaphorical competence in a second language (L2) is not static; it varies according to levels of familiarity and cross-linguistic correspondence (O'Reilly & Marsden, 2021). Furthermore, metaphor processing differs when comparing first-language and second-language contexts (Citron et al., 2020). In the context of BIPA (Indonesian for Foreign Speakers), cross-linguistic transfer has also been identified as a relevant dimension of Indonesian language learning (Permadi et al., 2022). The contribution

of this study lies in its juxtaposition of conceptual metaphor accuracy with verbally expressed strategy use, all synthesized within an exploratory comparison spanning four first-language (L1) backgrounds.

Within the context of BIPA research, this study addresses an under-researched aspect: how learners with diverse first language (L1) backgrounds interpret culturally conventional Indonesian figurative expressions. Previous research has indeed examined CEFR-based BIPA materials, learner discourse, instructional content, and teaching strategies (Fitria, 2023; Ningsih, 2023; Pratiwi et al., 2023; Subandiyah et al., 2023). From a psycholinguistic perspective, the findings of this study provide preliminary evidence that metaphor learning can benefit from taking L1-based conceptual resources into account, alongside factors such as Indonesian language proficiency and exposure to the language.

Theoretical, pedagogical, and linguistic-cultural implications

These results, when viewed from the standpoint of theory, underline the importance of a deeper understanding of cross-linguistic phenomena. The explanation offered here links the understanding of figurative language with the use of conceptual resources drawn from pre-existing linguistic experiences as well as from the maturation of Indonesian. The gathered data are consistent with the conceptual mapping theory. Still, the data cannot yet provide a clear picture of the distance or closeness between the first language (L1) as a whole and Indonesian. Only in case the analysis unit is composed of some definite conceptual or linguistic mappings or elements verified over a range of languages will we be able to make a clear-cut statement instead of being bound up with generalized language labels as the case is here.

Metaphors in the Indonesian language sit at the point where three different aspects of language meet: lexical form, well-established conceptual associations, and knowledge of the culture of the target audience. That position, both linguistically and culturally, makes metaphors a rich field of study. In BIPA classes, instructors can choose a contrastive approach as one of the methods. Students are asked to compare Indonesian metaphors with potential equivalents in their first language (L1). The goal is to determine whether the conceptual mappings they have are aligned. Direct equivalents do not always exist. However, meaning can still be inferred from the discourse context. Course materials should be organized progressively. Begin with mappings that are easily recognizable, then develop toward expressions that are culturally specific. Contextual scenarios should also be introduced. In this way, students become skilled at independently interpreting meaning in Indonesian.

These findings carry significant pedagogical implications, especially for learners who rely on translation services for almost all of their learning activities. It should be emphasized that translation should not be treated as an error. On the contrary, it can serve as an initial diagnostic tool to identify and assess learners' conceptual resources, while also providing a foundation for developing the ability to draw contextual conclusions, construct inferences, and

formulate interpretations that refer to cultural contexts. At the same time, it also reinforces interlingual awareness. The acquisition of figurative language skills in the instruction of Indonesian as a Foreign Language (BIPA) should also be considered an essential part of the full instructional cycle.

Research limitations and suggestions for future research

Several limitations define the scope of the findings. A total of 20 second language users were interviewed for this project, five coming from each L1 group, which is too few a number of cases to ensure reliable statistics and makes the group estimates highly dependent on individual performances. The study is therefore framed throughout as exploratory. Proficiency classification was based on the institutional BIPA placement assessment rather than on a separate study-specific standardized test, and the very small L1-by-proficiency cells precluded stable subgroup estimation. Moreover, duration, intensity, and quality of Indonesian exposure were not modeled independently, limiting the extent to which proficiency-related patterns can be separated from learning history.

A second limitation concerns the four L1 groups examined here: English, Mandarin, Japanese, and Arabic. Future research should include larger and more diverse samples, with balanced proficiency cells and explicit measures of length and intensity of Indonesian exposure. Additional L1 backgrounds, including languages with historically or regionally closer contact with Indonesian such as Malay and Thai, would permit more precise testing of mapping-level correspondence without treating languages as fixed points on a single conceptual-distance scale.

A third limitation concerns psychometric and statistical scope. The conceptual-equivalence classification was independently validated by two native-speaker validators for each L1, with high inter-rater agreement (κ = .80–.86), and 30% of the think-aloud transcripts were independently double-coded with high agreement (κ = .84). However, the inferential analysis aggregated responses across all 40 items and did not estimate a separate within-participant effect of equivalent versus non-equivalent metaphors. The archived analysis record also did not contain the item-level response matrix required to estimate Cronbach's alpha retrospectively, and formal Shapiro–Wilk and Levene test outputs were not retained. Future studies should retain complete item-level data and assumption diagnostics, preregister analysis decisions, test conceptual-equivalence categories directly, and model Indonesian exposure explicitly.

IV. CONCLUSION

In relation to RQ1, the findings indicate that Indonesian conceptual metaphor comprehension varied across the four L1 groups of BIPA learners. The English-L1 group showed the highest mean accuracy. Post-hoc comparisons showed significant differences between the English-L1 group and the Japanese- and Arabic-L1 groups. The remaining

pairwise contrasts were not statistically significant. In relation to RQ2, participant-level inspection suggested a descriptive proficiency-related pattern. C1 participants appeared to achieve higher comprehension accuracy than B1 participants in the available sample. However, no inferential proficiency effect or proficiency-by-L1 interaction was estimated.

Think-aloud coding revealed different interpretive strategies across the four L1 groups. Direct translation was more frequent in the Japanese- and Arabic-L1 groups, whereas contextual inference was more frequent in the English-L1 group. Overall, these findings are consistent with a possible role for L1 conceptual resources, Indonesian proficiency, item familiarity, and language exposure. However, this study has not yet disentangled these factors from one another.

This study provides preliminary Indonesian-language evidence contributing to cross-linguistic research on figurative-language comprehension and BIPA instruction. Pedagogically, teachers may use contrastive mapping, contextualized explanations, and guided inference to support metaphor comprehension. The difficulties faced by learners who still rely heavily on translation can be gradually addressed by focusing on Indonesian discourse and paraphrasing strategies. The conceptual-equivalence classification showed high agreement among the native-speaker validators across the four L1s ($\kappa = .80–.86$). However, the small sample, unavailable assumption-test outputs, and absence of the archived item-level response matrix require cautious interpretation of the findings. Future studies should use larger and more balanced samples, retain item-level and assumption-diagnostic data, test conceptual equivalence directly, and model Indonesian exposure explicitly.

DECLARATION

Author contribution: RN was involved in formulating the idea of the study, planning of research, data collection, data analysis, interpretation of results, writing of the manuscript, and critically revising the manuscript. HI contributed to the review of literature, data validation, interpretation of results, revising the manuscript, and also agreed to the final version submitted. Both authors read and approved the final manuscript and are responsible for all aspects of the work.
Funding statement: There was no funding source for this work from any public, commercial, or non-profit sectors.
Conflict of interest: The authors declare no conflict of interest related to the manuscript.
Additional information: No additional information is available for this paper.

REFERENCES

Boers, F., & Muñoz-Basols, J. (2021). *Acquisition of idiomatic language in L2 Spanish. In Spanish vocabulary learning in meaning-oriented instruction* (pp. 62–88). Routledge. https://doi.org/10.4324/9781315100364-5

Carston, R., & Yan, X. (2023). Metaphor processing: Referring and predicating. *Cognition*, 238, 105534. https://doi.org/10.1016/j.cognition.2023.105534

Cienki, A. (2022). The study of gesture in cognitive linguistics: How it could inform and inspire other research in cognitive science. *WIREs Cognitive Science, 13*(6), e1623. https://doi.org/10.1002/wcs.1623

Citron, F.M.M., Michaelis, N., & Goldberg, A.E. (2020). Metaphorical language processing and amygdala activation in L1 and L2. *Neuropsychologia*, 140, 107381. https://doi.org/10.1016/j.neuropsychologia.2020.107381

Febriansyah, G. (2020). Conceptual metaphor of anger emotion in Grunge Musician's song lyrics. *NOTION: Journal of Linguistics, Literature, and Culture*, *2*(1), 36–41. https://doi.org/10.12928/notion.v2i1.1026

Fitria, T.N. (2022). Figurative language in the Broken Wings by Kahlil Gibran: An analysis of language style as stylistic effect. *NOTION: Journal of Linguistics, Literature, and Culture*, *4*(2), 105–117. https://doi.org/10.12928/notion.v4i2.5798

Fitria, T.N. (2023). Common European framework of reference for languages (CEFR) standard in Bahasa Indonesia bagi penutur asing (BIPA) textbook: A competency mapping analysis. *Lingua Didaktika: Jurnal Bahasa dan Pembelajaran Bahasa, 17*(2), 139–156. https://doi.org/10.24036/ld.v17i2.124244

Gibbs, R.W. (2023). Pragmatic complexity in metaphor interpretation. *Cognition*, 237, 105455. https://doi.org/10.1016/j.cognition.2023.105455

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

MacArthur, F., & Alejo-González, R. (2024). Beyond idioms, the use of metaphor in ELF academic settings: A comprehensive review. *Journal of Pragmatics*, 219, 48–57. https://doi.org/10.1016/j.pragma.2023.11.002

Ningsih, H. K. (2023). Analisis genre dan APPRAISAL tulisan eksposisi pemelajar bahasa Indonesia bagi penutur asing. *Jurnal Bahasa Indonesia bagi Penutur Asing (JBIPA), 5*(2), 144–160. https://doi.org/10.26499/jbipa.v5i2.6709

O'Reilly, D., & Marsden, E. (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., & Marsden, E. (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

Permadi, R.Y., Nasrullah, R., & Koswara, I. (2022). The constraints of cross-linguistic transfer on BIPA learning. *Sosiohumaniora, 24*(2), 151–159. https://doi.org/10.24198/sosiohumaniora.v24i2.37363

Pratiwi, Y., Andajani, K., Suyitno, I., Ismail, A., & Prastio, B. (2023). Representing and implementing moral values to foreign students in Indonesian textbooks for learners other than Indonesians. *International Journal of Language Education, 7*(1), 58–76. https://doi.org/10.26858/ijole.v1i1.36256

Putri, P.B.Y., Arifin, M.A., Lestari, H.D., Nasrullah, R., & Rohim, R. (2025). Language and communication strategies for climate change engagement among Indonesian millennials and gen z. *NOTION: Journal of Linguistics, Literature, and Culture*, *7*(2), 332–348. https://doi.org/10.12928/notion.v7i2.12578

Rizzato, I. (2021). Conceptual conflicts in metaphors and pragmatic strategies for their translation. *Frontiers in Psychology*, 12, 662276. https://doi.org/10.3389/fpsyg.2021.662276

Subandiyah, H., Supratno, H., Permata, R., Rosyid, A., Minarti, E., & Nur Abida, F.I. (2023). Teaching Indonesian language for foreign speakers in the State University of Surabaya: Strategies and challenges. *Randwick International of Education and Linguistics Science Journal, 4*(2), 193–198. https://doi.org/10.47175/rielsj.v4i2.667

Tong, X., Shutova, E., & Lewis, M. (2021). Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In Proceedings of the 2021 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4673–4686). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.372

Yu, N., & Wang, B.P.-Y. (2019). Cognitive linguistics approaches to Chinese second language acquisition. In *The Routledge handbook of Chinese second language acquisition* (pp. 31–47). Routledge. https://doi.org/10.4324/9781315670706-2