frontiers in Language Sciences

TYPE Original Research
PUBLISHED 11 August 2026
DOI 10.3389/flang.2026.1872761



OPEN ACCESS

EDITED BY
Pedro Guijarro-Fuentes,
University of the Balearic Islands, Spain

REVIEWED BY
Agnès Lacroix,
University of Rennes 2 – Upper Brittany, France
Óscar Loureda,
Heidelberg University, Germany

*CORRESPONDENCE
Lena Wieland
✉ lena.wieland@uni-saarland.de

RECEIVED 05 May 2026
REVISED 17 June 2026
ACCEPTED 22 June 2026
PUBLISHED 11 August 2026

CITATION
Wieland L and Reich I (2026) Familiarity and literal plausibility, not transparency, drive idiom comprehension in adult readers with low literacy skills in Easy German. *Front. Lang. Sci.* 5:1872761. doi: 10.3389/flang.2026.1872761



# Familiarity and literal plausibility, not transparency, drive idiom comprehension in adult readers with low literacy skills in Easy German

Lena Wieland* and Ingo Reich

Department of German Studies, Saarland University, Saarbrücken, Germany

**Introduction:** Easy German (Leichte Sprache) is a rule-guided variety of German intended to increase the accessibility of written information for people with cognitive disabilities and adults with low literacy skills. Although existing guidelines often discourage the use of idioms, experimental evidence on idiom processing in these target groups remains scarce. We investigated how literacy, functional familiarity, transparency, and literal plausibility jointly shape idiom comprehension and preferences in adult low-literacy readers.

**Methods:** In a norming study with literate German speakers (Experiment 1; Prolific; 71 verb-phrase idioms), we obtained seven-point ratings of transparency and literal plausibility (reverse-coded so that higher values indicate a more plausible literal scene) and selected 24 idioms to instantiate a 2 × 2 design (high/low transparency × high/low plausibility). Literacy in the target group was assessed with otu.lea (Experiment 2) and transformed into a quasi-continuous index (0–100). In Experiment 3 (no context), 30 low-literacy adults explained each idiom, providing a measure of functional familiarity. In Experiment 4 (same cohort after ~6 months; $n = 26$), idioms were embedded in Easy German contexts and tested using an endorsement task (idiom vs. paraphrase vs. distractor) and a ranking task (gold/silver/bronze).

**Results:** Mixed-effects models showed that, in Experiment 3, figurative accuracy increased with literacy ($\beta = 0.028$, $p = 0.0019$) and was related to literal plausibility ($\beta = 0.569$, $p < 0.001$), with a strong transparency × plausibility interaction ($\beta = -1.457$, $p < 0.001$): idioms that were both transparent and literally plausible yielded the lowest figurative accuracy. In Experiment 4, idiom endorsement was predicted by functional familiarity from Experiment 3 ($\beta = 0.702$, $p = 0.025$), while ranking responses showed robust distractor rejection and no reliable disadvantage for idioms relative to meaning-consistent paraphrases.

**Discussion:** Overall, idiom accessibility in low-literacy adults appears to be driven primarily by literacy, functional familiarity, and literal-figurative competition indexed by plausibility, whereas transparency does not emerge as a reliable stand-alone facilitator. These findings challenge Easy German recommendations that prioritize transparency and instead support applied strategies centered on audience familiarity, low literal plausibility, and contextual or paraphrastic scaffolding.

KEYWORDS

**Easy German (Leichte Sprache), familiarity, figurative language, idiom comprehension, individual differences, literal plausibility, low literacy, transparency**

# 1 Introduction

Easy German (Leichte Sprache) is a strongly rule-governed variety of German intended to improve access to written information for people with cognitive disabilities and for adults with limited literacy skills (Bredel and Maaß, 2016; Bock and Pappert, 2023). In many applied contexts, these texts carry legal, medical, or civic consequences, for example in administrative correspondence, election information, patient instructions, and consent materials (Bredel and Maaß, 2016; DIN, 2025; Netzwerk Leichte Sprache, 2022). Because these consequences are tangible, the linguistic choices recommended by Easy German guidelines effectively function as hypotheses about comprehension (Bredel and Maaß, 2016; DIN, 2025). One recurring recommendation concerns figurative language. Several widely used rule sets advise avoiding figurative expressions, while others allow them under specific conditions, for example when an expression is common, when the context licenses the intended meaning, or when the idiom is considered transparent (Bredel and Maaß, 2016; DIN, 2025). Despite the practical importance of these recommendations, there is limited experimental evidence on whether low-literacy adults with and without cognitive impairments actually experience idioms as a systematic barrier, and on which idiom properties predict success or failure (cf. Nippold and Rudzinski, 1993; Nippold and Taylor, 2002; Cain et al., 2005; Levorato et al., 2004).

The present study was designed to address a concrete applied question, can adult low-literacy readers understand idioms in Easy German, and if so, which factors reliably facilitate or hinder comprehension. Rather than treating figurativity as a categorical risk, we focus on four factors that recur in both psycholinguistic accounts and Easy German rule discourse, literacy level, familiarity with a specific idiom, transparency (the extent to which figurative meaning can be inferred from literal components), and the plausibility of a literal interpretation. In psycholinguistic terms, these factors are often treated as cues that shape the relative activation and selection of literal and figurative interpretations (Swinney and Cutler, 1979; Gibbs et al., 1989; Titone and Connine, 1999; Titone and Libben, 2014). In applied terms, they correspond to concrete decisions about which expressions to include, which to avoid, and when to paraphrase or contextualize (Bredel and Maaß, 2016; DIN, 2025; Netzwerk Leichte Sprache, 2022).

Our research logic follows the structure of the questions faced by Easy German practitioners. Before asking whether an idiom is preferred or avoided, one must establish whether the target audience recognizes the idiom's figurative meaning at all. Similarly, before examining literacy-related individual differences, one must establish a stimulus set that systematically varies in item-level properties. We therefore implemented four interlocking studies.

Experiment 1 is a norming study conducted with literate German speakers. It quantifies transparency and literal plausibility for a broad pool of candidate verb phrase idioms and provides the empirical basis for selecting 24 idioms that instantiate a $2 \times 2$ design (transparency: high, low; literal plausibility: high, low). This step ensures that the manipulation of item-level properties is explicit and reproducible.

Experiment 2 assesses literacy in the target population using a low-threshold diagnostic [otu.lea, (Koppel and Wolf, 2014)]. Because literacy is expected to vary substantially within adult low-literacy groups, we treat it as a continuous individual-difference variable rather than a categorical label. This allows us to model how literacy modulates comprehension outcomes across tasks.

Experiment 3 then addresses the most basic question about idiom accessibility. Participants receive the 24 idioms without context and are asked to explain their meaning. This task is intentionally demanding because it requires participants to access a figurative interpretation and to externalize that interpretation in speech. We use this experiment as a no-context diagnostic to establish functional familiarity, that is, whether the figurative meaning is available to the participant in the absence of contextual scaffolding. At the same time, because the idioms were selected to vary in transparency and plausibility, Experiment 3 tests whether these item-level cues support or hinder the recovery of a figurative meaning under maximal competition conditions.

Experiment 4 moves from baseline access to idiom use under supportive conditions. The same participant cohort as in Experiment 3 returned after a ∼6-months interval (with four participants not re-tested for documented reasons), a delay chosen to reduce carryover and make it unlikely that item-specific responses were simply remembered from the earlier no-context session. Participants read short Easy German contexts designed to license the intended figurative meaning and then complete a multiple-choice comprehension task (idiom, meaning-consistent paraphrase, distractor). They then rank the options using an accessible podium metaphor (gold, silver, bronze). This design separates two issues that are often conflated in guideline discourse, comprehension and preference. An idiom may be understood but still avoided, or it may be acceptable once understood. Experiment 4 therefore allows us to assess whether idioms are dispreferred relative to literal paraphrases when meaning is available.

Across the four studies, we test the following hypotheses, which are intentionally formulated at the level of cue effects rather than at the level of specific processing architectures. First, literacy should positively predict idiom comprehension, especially in tasks that require self-generated interpretations. Second, familiarity should facilitate comprehension under contextual support, reflecting stored idiom knowledge. Third, transparency should facilitate comprehension in principle, but only to the extent that participants can exploit analytic cues from literal words. Fourth, literal plausibility should increase competition from literal interpretations and reduce figurative accuracy, particularly when context is absent. These hypotheses directly speak to Easy German practice, because transparency and familiarity are frequently invoked as licensing criteria, while literal plausibility is rarely discussed despite being a plausible source of literal bias.

In what follows, we first provide background on Easy German and on the psycholinguistics of idioms, then describe each experiment in turn, and finally return to the broader theoretical and practical implications.

# 2 Background

## 2.1 Easy German and figurative meaning

Easy German is often described in public discourse as "simplified German," but in practice it is better characterized as a bundle of constraints that jointly target decoding effort, working-memory load, and inferential burden. Rule sets such as Duden Leichte Sprache (Bredel and Maaß, 2016), the current DIN specification (DIN, 2025), and the Regelwerk of Netzwerk Leichte Sprache (Netzwerk Leichte Sprache, 2022) provide guidance on lexical choice, morphosyntax, information structure, and layout. Recommendations typically include short sentences with single propositions, avoidance of embedding, high-frequency vocabulary, explicit introduction of concepts, and typographic segmentation through line breaks and lists.

Example (1) Easy German sample text[1]

(1a) Leichte Sprache (German)

Dieser Text ist in Leichter Sprache geschrieben.
Leichte Sprache bedeutet:
Der Text ist einfach.
Jeder Satz hat nur eine Aussage.
Jeder Satz steht in einer Zeile.

(1b) Gloss (English)

This text is written in Easy German.
Easy German means:
The text is simple.
Each sentence has only one statement.
Each sentence is on its own line.

Figurative language sits uneasily within this regime. On the one hand, figurative expressions are pervasive in everyday German and can serve discourse goals such as evaluation, stance marking, and cohesion. On the other hand, figurative meaning often requires additional inference, and several guideline documents treat figurativity as an avoidable source of ambiguity–particularly in informational texts. Importantly, however, recommendations are not uniform. Some sources advocate avoiding figurative language altogether (Netzwerk Leichte Sprache, 2022). Others adopt a more differentiated stance, for instance by classifying metaphors or idioms according to their assumed impact on comprehension and allowing figurative expressions when they are familiar, inferable from context, or when they concretize otherwise abstract content. Where figurative language is central to a text, explicit explanation is sometimes recommended (Bredel and Maaß, 2016; DIN, 2025). This mixture of prohibition and conditional allowance implies implicit processing assumptions: that unfamiliar figurative expressions are disproportionately difficult, that context can compensate for limited lexicalized knowledge, and that transparency makes an expression "safer." These assumptions are plausible, but they are not trivial. Transparency, in particular, is an attractive accessibility heuristic because it suggests that meaning is partially "visible" in the words themselves; yet exploiting transparency may depend on analytic and metalinguistic resources that plausibly vary with literacy. The present paper treats these guideline claims as empirically testable hypotheses and aims to provide evidence that can inform future rule refinement.

The lack of consensus within Easy German guidance is consequential because it directly shapes institutional communication practices that affect participation, rights, and access to information. At the same time, the empirical evidence base on how Easy German texts, and especially potentially risky constructions such as figurative expressions, are processed by primary target audiences remains limited, and systematic experimental work is only beginning to emerge. One of the most influential applied research programs in this area is the LeiSa project (Leichte Sprache im Arbeitsleben). LeiSa explicitly challenges the idea that Easy German can be reduced to a monolithic set of prescriptive rules and instead examines how different target groups, including adults with intellectual disabilities, process and evaluate Easy German texts in realistic settings. Using comprehension tests, acceptability judgments, and qualitative interviews, LeiSa demonstrates that the effectiveness of Easy German depends strongly on text type, communicative setting, and reader background knowledge, and that the audience is markedly heterogeneous rather than approximating a single "ideal reader" profile (Bock, 2019). These findings motivate a shift away from purely prescriptive recommendations toward empirically grounded guidance and they motivate controlled psycholinguistic experimentation as a complementary approach: isolating which linguistic properties (and which task conditions) reliably support or hinder comprehension within heterogeneous readerships.

Our work also connects to recent German-language methodological contributions and infrastructures that formalize best practices for research with target groups of Leichte Sprache. This includes work associated with the "Simply Complex–Easy German" initiative in Mainz (e.g., Borghardt et al., 2021) and methodological reports from Universität Hildesheim (e.g., Rink et al., 2023), which emphasize recruitment via gatekeepers, accessibility-adapted task design, mediated response procedures, and transparent reporting of feasibility constraints. By situating our study alongside these applied and methodological initiatives, we aim to contribute to a converging evidence base in German accessibility research, one in which guideline claims, particularly those concerning figurative language, are evaluated against experimentally controlled data from the target audiences themselves.

## 2.2 Idioms as a test case for accessibility

Idioms are a particularly interesting case because they occupy a middle ground between lexicon and discourse (e.g., Cacciari and Tabossi, 1988; Wray, 2002). Many idioms are conventionalized, frequently used, and stored as multiword units, but they also remain syntactically productive to varying degrees (Cacciari and Tabossi, 1988; Gibbs et al., 1989). In the present study we focus on German verb phrase idioms, multiword expressions centered around a verb plus one or more particles or prepositions whose overall meaning is idiomatic and not fully predictable from the meanings of the parts (Donalies, 2012; Cap et al., 2018; Ehren et al., 2020). Examples include *jemandem die Daumen*

1 Source: German Federal Government, "Leichte Sprache". Accessed on 26 April 2026.

*drücken* "to wish someone luck" (lit. "to press the thumbs for someone"), *blau machen* "to skip work or school" (lit. "to make blue"), *an einem Strang ziehen* "to work together toward a common goal" (lit. "to pull on the same rope"), and *zittern wie Espenlaub* "to shake like a leaf" (lit. "to tremble like aspen leaves") (Donalies, 2012).

Following standard psycholinguistic characterizations, idioms can be defined as expressions whose meanings cannot be directly inferred from the meanings of their individual words (Glucksberg, 2001). At the same time, idioms vary systematically in transparency (Gibbs et al., 1989; Cacciari and Glucksberg, 1991; Vega-Moreno, 2007). In transparent idioms, aspects of figurative meaning can be related to the literal scene and its components, for example *am Ball bleiben* ("to keep at it," literally *to stay on the ball*) (Vega-Moreno, 2007). In opaque idioms, this mapping is less available, for example *jdm. auf der Nase herumtanzen* ("to walk all over someone," literally *to dance around on someone's nose*) (Gibbs et al., 1989; Vega-Moreno, 2007). Transparency is therefore often treated as a proxy for compositional support (Gibbs et al., 1989; Titone and Connine, 1999).

Idioms also vary in the plausibility of their literal interpretations (Titone and Connine, 1994; Cronk and Schweigert, 1992). Some idioms describe literal scenes that are odd or unlikely (low literal plausibility), for example *auf Wolke sieben sein* ("to be on cloud nine," literally *to be on cloud seven*), while others describe scenes that are easy to imagine and could, in principle, occur (high literal plausibility), for example *den Stier bei den Hörnern packen* ("take the bull by the horns," literally *to grab the bull by the horns*) (Titone and Connine, 1994). This distinction matters because literal plausibility is a natural source of competition (Titone and Connine, 1999; Libben and Titone, 2008). If a reader can easily construct a coherent literal scene, the literal candidate may remain attractive and may need to be inhibited or reinterpreted for a figurative meaning to prevail (Titone and Connine, 1999; Nordmann et al., 2014).

These item-level properties are not merely theoretical constructs. They align with intuitive practitioner concerns: "is the expression self-explanatory," "could it be misunderstood literally," and "do readers already know it," which correspond closely to transparency, literal plausibility (or literality), and familiarity as operationalized in idiom norming studies (Titone and Connine, 1994; Bonin et al., 2013; Morid and Sabourin, 2024). The present work aims to quantify these properties and relate them to comprehension outcomes in the target audience, following the logic of prior norming-based prediction studies (e.g., Libben and Titone, 2008; Tabossi et al., 2011).

## 2.3 Processing accounts and cue-based predictions

The idiom processing literature offers several families of accounts that differ in architecture and time course, but they converge on the idea that multiple constraints shape interpretation (Gibbs et al., 1989; Libben and Titone, 2008; Titone and Connine, 1999; Titone and Libben, 2014). Retrieval-based views propose that familiar idioms are stored as lexicalized units, allowing rapid access to figurative meaning without obligatory literal computation (Swinney and Cutler, 1979; Glucksberg, 2001). On this view, familiarity should be a dominant predictor, especially in tasks that provide enough cues to trigger retrieval (Titone and Connine, 1994; Libben and Titone, 2008).

Literal-first pragmatic accounts assume that literal meaning is computed and evaluated first, and that figurative meaning is derived when literal interpretation fails to achieve coherence or relevance (Grice, 1975; Searle, 1979; Glucksberg, 2001). Here, literal plausibility should be central, because a plausible literal interpretation will delay or block the shift to figurative meaning (Titone and Connine, 1999; Tabossi et al., 2011).

Configurational and decomposition-based accounts propose that readers process idioms incrementally and that transparency or analyzability facilitates comprehension once the idiom is recognized (Gibbs et al., 1989; Cacciari and Glucksberg, 1991). Under these accounts, transparency should be beneficial, particularly when an idiom is unfamiliar and retrieval is weak (Gibbs et al., 1989; Nordmann et al., 2014).

Constraint-based competition frameworks aim to integrate these views by treating comprehension as the result of parallel activation and competition among interpretations, with outcomes determined by the weighted integration of probabilistic cues (Titone and Connine, 1999; Libben and Titone, 2008; Titone and Libben, 2014). In such models, familiarity increases support for the figurative candidate, plausibility increases support for the literal candidate, and transparency can support either candidate depending on whether the reader can execute an analytic mapping from literal components to figurative meaning (Titone and Connine, 1999; Libben and Titone, 2008).

Crucially for the present study, these models are largely developed from highly literate samples (Perfetti, 1985; Nippold, 1998; Titone and Libben, 2014). Literacy may modulate cue use in at least three ways. First, low literacy can limit lexical and semantic knowledge, reducing access to stored idiom meanings (Perfetti, 1985; Nation, 2005). Second, low literacy can constrain executive control and metalinguistic awareness, reducing the ability to maintain and compare alternatives (Nippold, 1991; Cain et al., 2005). Third, low literacy can increase the cost of reading itself, leaving fewer resources for inferential processing (Perfetti, 1985; Nation, 2005). These considerations motivate treating literacy as a continuous predictor that may modulate cue weights, in line with broader individual-differences work in figurative and idiom processing (Nippold, 1998; Libben and Titone, 2008).

The study's hypotheses translate these theoretical considerations into concrete predictions about the four focal factors. Literacy and familiarity are expected to facilitate comprehension, transparency is expected to help when analytic mapping is available, and literal plausibility is expected to hinder comprehension by strengthening the literal competitor (Gibbs et al., 1989; Titone and Connine, 1999; Titone and Libben, 2014; Nippold, 1991).

## 2.4 Hypotheses

Based on the background above and the practical requirements of Easy German communication, we tested four hypotheses.

H1. Literacy. Higher literacy scores predict higher idiom comprehension accuracy.

H2. Familiarity. Familiarity with an idiom increases comprehension accuracy, particularly under contextualized comprehension demands.

H3. Transparency. Transparent idioms, whose figurative meanings can be inferred from their literal meanings, are easier to interpret than opaque idioms.

H4. Literal plausibility. High literal plausibility increases competition with figurative interpretations and reduces accuracy, especially when contextual support is absent.

# 3 Experiment 1, norming transparency and literal plausibility

## 3.1 Rationale

Experiment 1 served two purposes. First, it provided quantitative transparency and plausibility ratings for a large candidate pool of German verb phrase idioms. Second, it allowed us to derive a controlled 2 × 2 factorial stimuli set for the main experiments in the low-literacy target group.

## 3.2 Materials

### 3.2.1 Stimulus pool

Candidate expressions were drawn from an initial pool of approximately 2,500 German idiomatic expressions compiled from dictionaries, corpora resources, and investigator expertise. To maximize experimental control, we restricted the pool to conventionalized verb phrase idioms (Donalies, 2012). Items were checked against *Duden Redewendungen* (Dudenredaktion, 2021) to ensure contemporary usage. From this pool, 71 idioms were selected for norming. Selection aimed to cover a broad range of semantic domains, degrees of transparency, and degrees of literal plausibility.

### 3.2.2 List construction

To keep the task duration feasible for online participants, the 71 idioms were divided into two lists. Each list contained a comparable mix of idioms, and each was completed by a separate group of participants ($n = 30$ per list).

## 3.3 Participants

A total of 60 native speakers of German were recruited via Prolific (Palan and Schitter, 2018), with 30 participants per list. Each participant completed one list and received monetary compensation in line with the German statutory minimum wage at the time of data collection (€12 per hour), consistent with Prolific's payment guidelines. Unlike the low-literacy sample tested in the later experiments, participants in Experiment 1 were not selected for reading ability, as the aim was to obtain stable item-level norms from a general adult sample. Each trial presented a single idiom (verb-phrase idiom; Donalies, 2012).

## 3.4 Procedure and measures

Experiment 1 was implemented as an online survey in PennController for Ibex (PCIbex; Zehr and Schwarz, 2018). Each trial presented a single idiom. Participants first completed a brief production prompt intended to encourage engagement with meaning, and then provided ratings on seven-point Likert scales.

### 3.4.1 Production prompt

Participants first indicated whether they had heard the expression before (yes/no). They then provided a short paraphrase of what they believed it meant and rated how confident they were in that interpretation on a seven-point scale ranging from *completely sure* (1) to *completely unsure* (7). This component served as (i) an engagement check (ensuring participants attempted to access meaning rather than rating surface form), and (ii) an informal familiarity measure that complemented the later norming variables. Because paraphrases can be noisy and the correct interpretation is sometimes underspecified without context, paraphrase responses were evaluated by two independent annotators using a binary coding scheme (0 = incorrect/unclear, 1 = correct) against a predefined target meaning list. Inter-annotator agreement was quantified using Cohen's $\kappa$ (Cohen, 1960) and was moderate [$\kappa = 0.47$; $N = 2{,}130$ paraphrases; Landis and Koch, 1977]. Remaining disagreements were resolved through discussion between the first and last authors before the codes entered analysis.

### 3.4.2 Transparency rating

Participants rated how easily the idiom's figurative meaning could be inferred from the literal meanings of its component words if someone were not already familiar with the idiom. Ratings were collected on a seven-point Likert scale from *totally hard* (1) to *totally easy* (7). This operationalizes transparency as perceived inferability (Gibbs et al., 1989; Glucksberg, 2001).

### 3.4.3 Literal plausibility rating (reverse-coded to literal plausibility)

Participants rated how likely the literal scenario suggested by the idiom would be as a real-world event. To make the question accessible, we framed it as likelihood rather than logical possibility. Ratings were collected on a seven-point scale from *completely likely* (1) to *completely unlikely* (7). Because this scale increases with unlikeliness, we reverse-coded it for analysis so that higher values indicate higher literal plausibility (i.e., a more plausible literal scene):

$$\text{Literal plausibility} = 8 - \text{Unlikeliness rating}$$

All analyses and plots in this manuscript use the reverse-coded metric unless otherwise stated. All items were presented in randomized order within each list.

## 3.5 Results

Ratings were aggregated by item. For each idiom, we computed mean transparency and mean literal plausibility across raters. In addition to means, we inspected item-level dispersion (standard deviations and distribution shape) to identify items that elicited unstable or strongly bimodal judgments. Items with unusually high dispersion were treated as poorer candidates for categorical grouping because high variance suggests the item does not reliably instantiate a "high" or "low" level on the intended dimension.

Figure 1 visualizes the distribution of the aggregated ratings for the final item set used in the main experiments. The plot makes two points visible. First, the selected items spanned a broad range on both dimensions: mean transparency ranged from 2.63 to 5.87 ($M$ = 4.22) and mean literal plausibility from 1.63 to 5.70 ($M$ = 3.27). Within each dimension, the high and low subsets occupied clearly separated portions of the scale, with no overlap between categories: low-transparency items ranged from 2.63 to 3.87 and high-transparency items from 4.63 to 5.87, while low-plausibility items ranged from 1.63 to 2.73 and high-plausibility items from 3.93 to 5.70. Second, transparency and plausibility were not perfectly coupled in the selected set, which supported constructing a factorial design in which the two dimensions could be manipulated independently at the level of item groups.

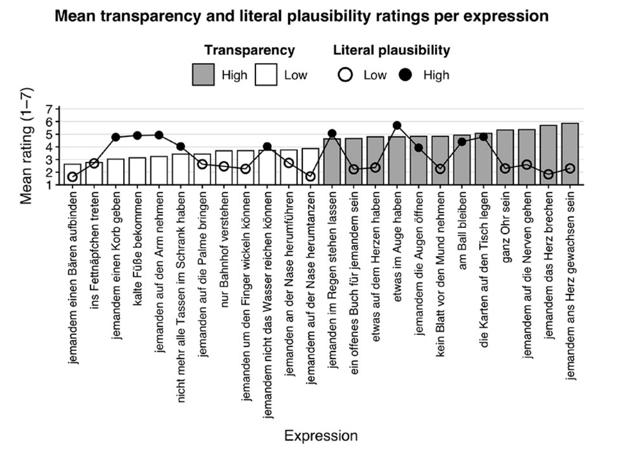


FIGURE 1
Norming study results for transparency and literal plausibility (final stimulus set). Bars show mean transparency ratings for each idiom (1–7; higher values indicate higher perceived transparency). The black line with points shows mean literal plausibility for the same items (reverse-coded so that higher values indicate a more plausible literal scenario), with point type indicating the dichotomized plausibility category (low vs. high). Items are ordered by mean transparency to illustrate the spread of values and the separation of the high/low category assignments.

## 3.6 Deriving the 2 × 2 design for Experiments 3 and 4

The goal of the norming study was to derive a stimulus set for Experiments 3 and 4 that cleanly instantiated a 2 × 2 manipulation of transparency, high vs low, and literal plausibility, high vs low, with the two dimensions crossed orthogonally. To maximize the separation between high and low categories while keeping item counts orthogonal, we used a trimming procedure based on ranked mean ratings. For transparency and literal plausibility, items were ranked by their mean rating on the respective dimension and split into three bands, low, middle, high. The middle band was excluded, because removing intermediate items increases the expected discriminability between high and low groups and reduces ambiguity in categorical assignment. From the remaining low and high candidate pools on each dimension, we selected 24 idioms that populated the four cells of the 2 × 2 design. Selection was guided by three constraints: (1), separation, items in the high and low categories were chosen to maximize the difference in mean ratings between groups, while avoiding items close to the band boundaries. (2), balance, the high- and low-plausibility proportions were held constant across transparency levels (a 7/5/7/5 distribution), ensuring that transparency and plausibility were not confounded with each other or with list membership. (3), stability, when multiple candidates were available, items with lower rating dispersion were preferred over items with higher dispersion, to reduce within-cell heterogeneity and to strengthen the interpretability of category-level contrasts. The resulting 24-item stimulus set, including item-level mean ratings and their dichotomized category assignments, is reported in Table 1.

TABLE 1 2 × 2 stimulus design crossing transparency (high, low) and literal plausibility (high, low) with example German idioms and glosses.

| | | Plausibility | |
|---|---|---|---|
| | | High | Low |
| Transparency | High | **am Ball bleiben** (lit.: stay on the ball, fig.: keep at it) | **etwas auf dem Herzen haben** (lit.: to have something on one's heart, fig.: to have something weighing on one's mind) |
| | Low | **jemanden auf den Arm nehmen** (lit.: take someone on the arm, fig.: tease or trick someone) | **jemandem auf der Nase herumtanzen** (lit.: dance around on someone's nose, fig.: to walk all over someone) |

Bold indicates the German idiom; literal (lit.) and figurative (fig.) glosses are given in parentheses.

TABLE 2 Fixed effects from the generalized linear mixed-effects model predicting accuracy in Experiment 3 (binomial logit).

| Predictor | $\beta$ (estimate) | SE | $z$ | $p$-value |
|---|---|---|---|---|
| (Intercept) | −1.4001 | 0.5939 | −2.358 | 0.0184 |
| Literacy score | 0.0275 | 0.00884 | 3.112 | 0.00186 |
| Transparency (sum-coded) | −0.1775 | 0.1663 | −1.068 | 0.2857 |
| Literal plausibility (sum-coded) | 0.5693 | 0.1669 | 3.411 | 0.000648 |
| Transparency × plausibility | −1.4571 | 0.3348 | −4.353 | <0.001 |

TABLE 3 Fixed effects from the mixed-effects logistic regression predicting idiom selection (Experiment 4).

| Predictor | $\beta$ | SE | $z$ | $p$-value |
|---|---|---|---|---|
| (Intercept) | 2.4158 | 0.3182 | 7.592 | <0.001 |
| Literacy (`score_c`) | 0.4992 | 0.2688 | 1.857 | 0.063 |
| Functional familiarity (Exp3 `accuracy_sum`) | 0.7019 | 0.3138 | 2.236 | 0.025 |
| Transparency (`tran_mean_c`) | 0.3333 | 0.1893 | 1.761 | 0.078 |

Transparency and plausibility were treated as categorical predictors in the main experiments to match the design logic and to reduce estimation instability in the small item set. At the same time, the underlying continuous ratings remain available for exploratory analyses and replication.

# 4 Experiment 2, literacy assessment and continuous literacy score

## 4.1 Rationale

Literacy is a defining characteristic of the intended audience of Easy German and a plausible moderator of figurative-language comprehension. We therefore assessed participants' reading-related skills prior to the idiom experiments and derived a continuous literacy score to be used as a participant-level predictor in Experiments 3 and 4. The goal was not to diagnose individuals clinically, but to capture variability within the target group that may translate into differences in lexical access, integration, and the ability to construct and maintain alternative interpretations.

## 4.2 Materials and procedure

Literacy was assessed with the otu.lea battery (Koppel and Wolf, 2014), a low-threshold, adaptive diagnostic widely used in adult-literacy instruction. The test yields an "alpha-level" classification that reflects performance across subcomponents spanning word-level decoding, sentence-level comprehension, and short-text understanding. The otu.lea battery samples functional reading competence across three task domains: word reading, sentence reading, and text reading by assigning participants to alpha levels that reflect the highest level of reading comprehension they can reliably demonstrate. In our implementation, the relevant competence descriptors were as follows.

### 4.2.1 Alpha level 3 (word reading/sentence reading)

Level 3 extends decoding beyond short word forms and introduces sentence-based comprehension: (3_01) decoding of words with more than five graphemes; (3_02) reading single words in sentence context; (3_03) sentence–picture matching; (3_04) reading sentences without insertions; (3_05) reading sentences with insertions; (3_06) following simple instructions, particularly when supported by pictures; and (3_07) reading TV schedules including time information.

### 4.2.2 Alpha level 4 (sentence reading/text reading)

Level 4 indexes the ability to navigate structured written materials and to extract both explicit and implicit information from short connected texts: (4_01) identifying and reproducing individual words from a text; (4_02) recognizing structures of simple forms; (4_03) extracting explicitly stated information from short texts (≤5 sentences; optionally supported by pictures/illustrations); and (4_04) deriving implicitly conveyed information from short texts (≤5 sentences; optionally supported by pictures/illustrations).

### 4.2.3 Alpha level 5 (text reading)

Level 5 targets comprehension of longer discourse and the recovery of explicit vs. implicit content beyond short-text contexts: (5_01) reading longer texts (>5 sentences) for meaning; (5_02) identifying and reproducing explicitly stated information from longer texts; and (5_03) identifying and reproducing implicitly conveyed information from longer texts.

Testing took place on site at the sheltered workshop (WfbM) of our cooperation partner *AWO Teilhabe Saarland*. Sessions were conducted individually in a quiet room provided by the facility to minimize distractions and ensure accessibility. A trained experimenter administered the battery in an accessibility-oriented way: participants responded orally and/or by pointing, and the experimenter entered responses. Where recommended by institutional practice, social workers provided an initial estimate of the participant's likely level so that the adaptive test could start at an appropriate difficulty range, reducing frustration and floor effects.

Participants received compensation aligned with the German minimum wage at the time of testing (approximately €12/h). We recruited 33 adult native speakers of German from the workshop. Two individuals, although initially screened as eligible

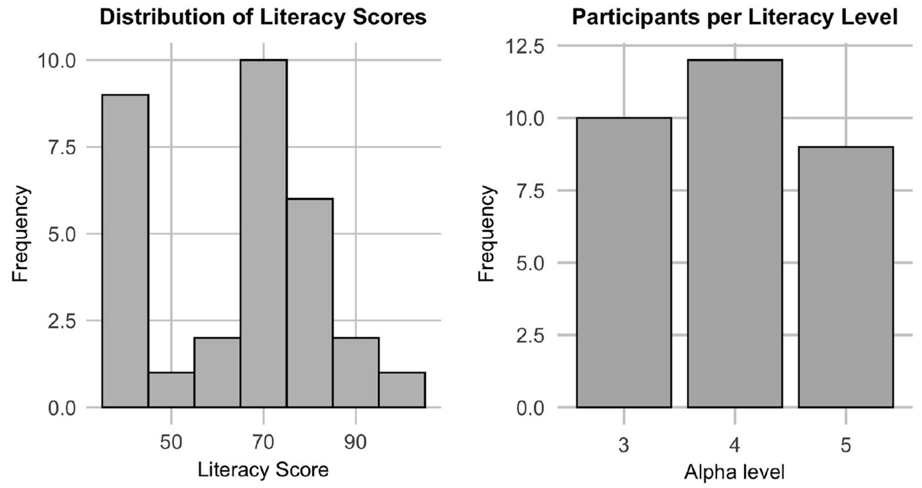


FIGURE 2
Literacy distribution in the target sample (otu.lea). **(Left)** distribution of participants' quasi-continuous literacy scores (0–100) derived from otu.lea performance (used as the continuous literacy predictor in Experiments 3 and 4). **(Right)** number of participants per otu.lea alpha level (Levels 3–5), showing the categorical classification on which the continuous index is based. Higher alpha levels reflect higher reading competence in the otu.lea framework (word-, sentence-, and text-level tasks).

by workshop staff, could not complete the tasks independently at the required basic reading level, making their data incomparable to the remaining sample; they were therefore not retained. Of the 31 participants who completed the literacy assessment and met the basic reading criterion, one was excluded for pre-specified safeguarding reasons, as the session could not be completed in a manner consistent with ethical and procedural requirements (persistent off-task content and safeguarding considerations). The final sample entering subsequent analyses therefore comprised 30 low-literacy readers. The included sample covered a range of alpha levels, consistent with heterogeneity that is typical in institutional settings.

## 4.3 Deriving a continuous literacy score

The raw output of otu.lea is categorical (alpha levels) and thus too coarse for individual-differences modeling. To preserve gradience while keeping the assessment accessible, we converted the categorical output into a quasi-continuous literacy index. Following the logic of the level structure, each attainable proficiency level (Levels 2–5) was assigned an incremental weight of 0.25. For each participant, performance within the highest completed level (proportion correct) was multiplied by that level's weight and added to the summed weights of all fully completed lower levels. The resulting value was rescaled to range from 0 to 100. This procedure retains information about both the level reached and accuracy within levels, and it provides a continuous predictor that can be entered into mixed-effects models.

To illustrate the procedure, consider a participant who fully completes Level 2 and reaches Level 3 with 80% correct at Level 3. Their score is computed as $100 \times (0.25 + 0.25 \times 0.8) = 45.0$. A participant who fully completes Levels 2 and 3 and reaches Level 4 with 60% correct at Level 4 receives $100 \times (0.50 + 0.25 \times 0.60) = 65.0$. Thus, the index jointly reflects how far a participant progresses in the hierarchy and how accurately they perform within their highest level.

Figure 2 shows the distribution of the resulting continuous literacy scores in the final sample. The resulting literacy scores ranged from 37.50 to 95.50 (median 70.00, mean 64.95; first quartile 44.75, third quartile 75.88). In all inferential analyses, literacy was treated as a continuous predictor; when literacy subgroups are shown in figures, this is for visualization only.

# 5 Experiment 3, baseline idiom access and functional familiarity

## 5.1 Rationale and design

Experiment 3 was designed as a no-context diagnostic of baseline idiom access in the target group. Participants received the 24 selected idioms without context and were asked to explain what each expression means. This task serves two functions. First, it provides a measure of whether a participant can access the idiom's figurative meaning without contextual scaffolding, which we treat as functional familiarity. Second, because idioms were selected to vary in transparency and literal plausibility, the task allows us to test whether these item-level cues predict success or failure when literal–figurative competition is maximal. The design therefore crosses two item-level factors (transparency and literal plausibility, each high vs. low) and includes literacy as a continuous participant-level predictor. This step therefore establishes the familiarity baseline for the subsequent contextualized comprehension and preference task (Experiment 4), where idioms are embedded in Easy German contexts.

Thirty adult native speakers of German from the same cohort as Experiment 2 (low-literacy readers) completed Experiment 3. Testing was conducted individually in a quiet room at the sheltered workshop (WfbM) of our cooperation partner *AWO Teilhabe*

*Saarland*, with experimenter-mediated responding to minimize device-related barriers. Participants received compensation aligned with the German minimum wage at the time of testing (approximately €12 per hour).

## 5.2 Materials and procedure

### 5.2.1 Stimuli

The stimulus set comprised 24 verb-phrase idioms [2 × 2 design: Transparency (high vs. low) × Literal plausibility (high vs. low); 7/5/7/5 distribution across cells]. Idioms were presented without context to probe baseline access to figurative meaning. Transparency and literal plausibility categories were derived from Experiment 1 (norming study) and treated as categorical predictors here.

### 5.2.2 Procedure

The experiment was administered individually and in person at the sheltered workshop. The task prompt was delivered in Easy German and presented both orally and in written form. Participants were asked, *Was heißt der Satz?* ("What does the sentence mean?"), and provided oral responses. The experimenter recorded responses verbatim whenever possible and read them back to the participant for confirmation. No corrective or evaluative feedback was given during the task.

### 5.2.3 Accessibility adaptations

The interface used large fonts and high contrast. Participants did not need to type. All responses were mediated by the experimenter to reduce confounds from device familiarity and motor demands.

## 5.3 Coding and dependent measure

Responses were transcribed and coded into three categories, figurative interpretation, literal interpretation, and uncodable or no interpretation. Figurative interpretations conveyed the conventional idiomatic meaning, either through paraphrase or through an example that clearly instantiated the intended meaning. Literal interpretations described the literal scene suggested by the words. Uncodable responses included refusals, unrelated answers, or insufficient content. For inferential analyses, we defined a binary accuracy measure: correct = accurate figurative interpretation (1) vs. incorrect = literal + unfamiliar/uncodable (0). This operationalization directly captures the idea that accuracy in the no-context task indexes functional familiarity/availability of idiomatic meaning. Three trained raters coded all responses using a shared codebook. Remaining disagreements were resolved through discussion between the first and last authors. The primary dependent variable for modeling was binary accuracy; figurative interpretation coded as 1 and all other responses coded as 0.

## 5.4 Statistical analyses

### 5.4.1 Descriptive results

Across the full dataset (30 participants × 24 idioms = 720 trials), responses were predominantly figurative. Of all trials, 550 (76.4%) were coded as figurative interpretations, compared to 60 literal interpretations (8.3%) and 110 "unfamiliar/uncodable" responses (15.3%; i.e., explicit don't-know responses, refusals, or answers that could not be mapped to either reading). Collapsing these categories into the preregistered binary outcome (accurate figurative interpretation = correct; figurative responses conveying an incorrect idiomatic meaning, literal, and unfamiliar/uncodable responses = incorrect), participants produced 428 correct responses (59.4%) and 292 incorrect responses (40.6%). Taken together, this pattern suggests that a sizeable subset of the stimulus set was functionally familiar to the target group even without contextual scaffolding, while the remaining errors, both literal construals and "unknown" responses, indicate substantial between-item and between-participant variability. This mix is exactly what one would expect under a graded familiarity view, where stored idiom knowledge varies in strength across expressions and where literal competition remains viable for some items in the absence of context.

As a feasibility check, we also summarized response length and testing time. Verbal productions varied widely in length (word count: min = 0, median = 6, mean = 9.05, max = 51), consistent with heterogeneity in expressive resources and response strategies in this population. Total task duration likewise showed considerable spread but remained practical for the workshop setting (minutes: min = 6.72, median = 21.55, mean = 21.94, max = 44.88).

### 5.4.2 Inferential analysis

To quantify which cues predict baseline access to idiomatic meaning in the absence of contextual support, we analyzed trial-level accuracy in Experiment 3 with a generalized linear mixed-effects model (binomial family, logit link) fitted with `lme4` (Bates et al., 2015) in R (R Core Team, 2023). The dependent measure was accuracy, coded as 1 when a response expressed the conventional figurative meaning and 0 otherwise (literal interpretations and unfamiliar/uncodable responses collapsed into the incorrect category). Fixed effects comprised (i) literacy as a continuous individual-difference predictor, (ii) transparency (high vs. low), (iii) literal plausibility (high vs. low), and (iv) the transparency × plausibility interaction. The two categorical predictors were sum-coded ($-0.5/+0.5$),[2] so that each main effect can be interpreted as an average effect across the levels of the other factor, rather than relative to an arbitrary reference level, and remains meaningful in the presence of the interaction. We included a by-participant random intercept to account for stable between-participant differences in baseline performance. By-item random

2 These predictors appear as `tran_cat_binary` and `plaus_cat_binary` in the analysis script; despite the "binary" label they were entered as sum-coded ($-0.5/+0.5$) contrasts.

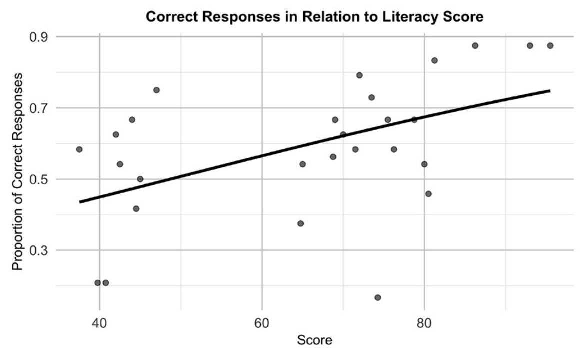


FIGURE 3
Literacy predicts idiom familiarity/interpretation accuracy in Experiment 3. Scatterplot of participant-level accuracy in the no-context production/interpretation task as a function of the continuous literacy score derived from otu.lea. Each point represents one participant's proportion of idioms correctly interpreted as figurative (correct figurative interpretations/24 items). The solid line shows the fitted relationship from the mixed-effects model, indicating higher literacy scores are associated with higher accuracy ($\beta$ = 0.0275, SE = 0.00884, $z$ = 3.11, $p$ = 0.00186).

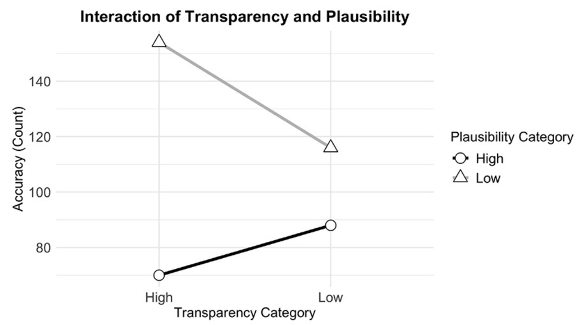


FIGURE 4
Interaction of transparency and literal plausibility in Experiment 3. Accuracy as a function of transparency category (high vs. low), plotted separately for literal plausibility category (high vs. low). The pattern illustrates a robust interaction ($\beta$ = −1.4571, SE = 0.3348, $z$ = −4.35, $p$ = 1.34 × $10^{-5}$), with the lowest figurative accuracy for idioms that are both highly transparent and highly literally plausible.

effects were not retained in the final model owing to convergence constraints in the small item set; as a robustness check, we inspected descriptive patterns by item and confirmed that the central effects were not driven by any single outlier idiom. The model was fitted to 720 trials (30 participants × 24 items), with an Akaike information criterion (AIC) of 901.3, a Bayesian information criterion (BIC) of 928.7, and a log-likelihood of −444.6.

## 5.5 Results

### 5.5.1 Random effects

Participants differed substantially in baseline performance: the random-intercept variance for id was 0.4236 (SD = 0.6509), indicating meaningful between-participant variability in overall likelihood of producing a correct figurative interpretation.

### 5.5.2 Literacy

Literacy reliably predicted performance. Higher literacy scores were associated with higher odds of producing a correct figurative interpretation ($\beta$ = 0.0275, SE = 0.00884, $z$ = 3.11, $p$ = 0.00186; Table 2). Figure 3 visualizes this participant-level pattern, showing a positive association between literacy and the proportion of idioms interpreted figuratively.

### 5.5.3 Literal plausibility and transparency

There was no overall main effect of transparency ($\beta$ = −0.1775, SE = 0.1663, $z$ = −1.07, $p$ = 0.286). Literal plausibility showed a significant main effect ($\beta$ = 0.5693, SE = 0.1669, $z$ = 3.41, $p$ = 0.000648). Because predictors were sum-coded and the model contained a strong interaction (see below), this main effect should be interpreted as an average difference across transparency levels and is best understood in conjunction with the interaction and the corresponding cell means.

### 5.5.4 Transparency × plausibility interaction

Critically, transparency interacted robustly with literal plausibility ($\beta$ = −1.4571, SE = 0.3348, $z$ = −4.35, $p$ = 1.34 × $10^{-5}$). As shown descriptively in Figure 4, figurative accuracy was lowest for idioms that combined high transparency with high literal plausibility, consistent with increased competition from an available literal construal in items where both compositional coherence and a plausible literal scenario jointly support a literal interpretation.

## 5.6 Interim discussion

Experiment 3 provides clear evidence that literacy matters for baseline idiom access in a no-context setting: participants with higher literacy scores were more likely to produce a correct figurative interpretation. This pattern is expected if literacy-related resources (e.g., lexical knowledge, integration, and control processes required to maintain and select among competing interpretations) support retrieval and articulation of idiomatic meanings when no contextual scaffolding is available. Beyond literacy, the results indicate systematic competition between literal and figurative candidates. Literal plausibility significantly modulated accuracy, but its contribution cannot be interpreted independently because it entered into a robust interaction with transparency. Transparency did not yield a uniform facilitation effect. Instead, the lowest figurative accuracy occurred for idioms that were simultaneously high in transparency and high in literal plausibility (Figure 4). One interpretation is that transparency increases the coherence of the literal scene, and when that scene is already plausible, the literal candidate becomes especially strong. If the analytic mapping from literal to figurative meaning is not reliably available, transparency can therefore amplify competition rather than resolve it. At the same time, the interaction suggests that transparency may still be beneficial under low plausibility, where the literal candidate is weak. In other words, transparency may help when it supports figurative inference more than it supports a literal construal. This is compatible with a conditional view of transparency. Because this interaction is unexpected in its direction relative to a straightforward 'transparent is easier' heuristic, it motivates two follow-up steps. First, we inspected item categorization to ensure that transparency and plausibility assignments match the norming data and the intended operationalizations. Second, Experiment 4 provides a complementary test, because it examines comprehension and preference under contextual support, where analytic resources may be less taxed and where choice formats reduce expressive demands.

# 6 Experiment 4, contextualized comprehension and preferences

## 6.1 Rationale and design

Experiment 4 moves from baseline access (Experiment 3) to a situation closer to Easy German practice. Idioms in real texts are typically embedded in supportive contexts, and writers often provide redundancy (e.g., an explicit paraphrase) to reduce the risk of misinterpretation. Our goal was therefore twofold: (i) to test whether participants endorse idioms as context-appropriate options when an Easy German context licenses the figurative meaning, and (ii) to test preferences between idioms, literal paraphrases, and distractors once these alternatives are jointly available. Crucially, Experiment 4 was designed as a delayed follow-up to Experiment 3 in order to minimize carryover effects from the no-context production task. The two sessions were separated by approximately 6 months, which reduces the likelihood that participants' responses in Experiment 4 reflect short-term item memory, repetition-based priming, or task-specific strategies acquired in Experiment 3, rather than context-supported interpretation under more naturalistic conditions.

First, a multiple-choice comprehension task assesses whether participants can recognize meaning in context when the idiom competes with a literal paraphrase and a distractor. Second, a ranking task assesses preference, that is, whether participants favor the idiom, the literal paraphrase, or the distractor once meaning is available. This distinction is important because guideline debates often conflate comprehension with acceptability.

The design uses the same 24 verb-phrase idioms and the same transparency × plausibility categorization as Experiment 3. Literacy remains an individual-difference predictor. Functional familiarity was derived from Experiment 3 and coded at the participant × item level (1 = correct figurative interpretation in Experiment 3; 0 = incorrect/unknown). An idiom was treated as functionally familiar for a given participant if that participant produced the correct figurative interpretation without context in Experiment 3. This provides a conservative, behavior-based estimate of whether idiomatic meaning was available prior to contextual scaffolding. In Experiment 4, functional familiarity was included as a key predictor and as a criterion for conditional preference analyses (ranking given comprehension).

Experiment 4 was conducted with the same cohort of adult native speakers of German recruited from the sheltered workshop (WfbM) described for Experiment 3. Of the 30 participants tested in Experiment 3, 26 completed Experiment 4. Attrition ($n = 4$) was due to practical constraints typical for longitudinal work in workshop settings (e.g., changes in availability, scheduling constraints, and health-related absences). No additional participants were recruited for Experiment 4. Testing was conducted individually in a quiet room at the workshop, and responses were mediated by the experimenter to minimize device-related barriers. Participants received compensation aligned with the German minimum wage at the time of testing (approximately €12 per hour).

## 6.2 Materials

### 6.2.1 Contexts

For each idiom, we created short Easy German contexts. Contexts were concrete and event-based, with explicit protagonists and causal relations. The goal was to license the intended figurative meaning without introducing additional lexical or syntactic complexity. To ensure that contexts were comprehensible for the intended audience, they were independently evaluated by an additional group of readers from the same target population (adult low-literacy readers). Participants in this check completed a brief comprehension screening (e.g., paraphrasing the situation or answering simple content questions) and indicated whether the context was understandable. Contexts that elicited recurrent misunderstandings were revised iteratively (simplified wording, clarified causal links) until they were reliably understood.

### 6.2.2 Options and interface

After the context, participants were presented with three response options: (i) the idiom embedded in a short declarative sentence, (ii) a literal paraphrase expressing the intended meaning without figurative language, and (iii) a distractor paraphrase unrelated to the intended meaning but matched for length and simplicity. Options were displayed in separate, spatially stable boxes to reduce visual search demands and to support pointing responses (Figure 5). Items were presented in a pseudo-randomized order, distributed across six lists, with participants assigned randomly and in equal numbers to the six lists. Within each list, the spatial arrangement of the three option types was held fixed per item but counterbalanced across items, so that no option type (idiom, paraphrase, or distractor) was consistently associated with a given screen position; this controls for position-based response biases while keeping the display stable for any given trial (Figure 5).

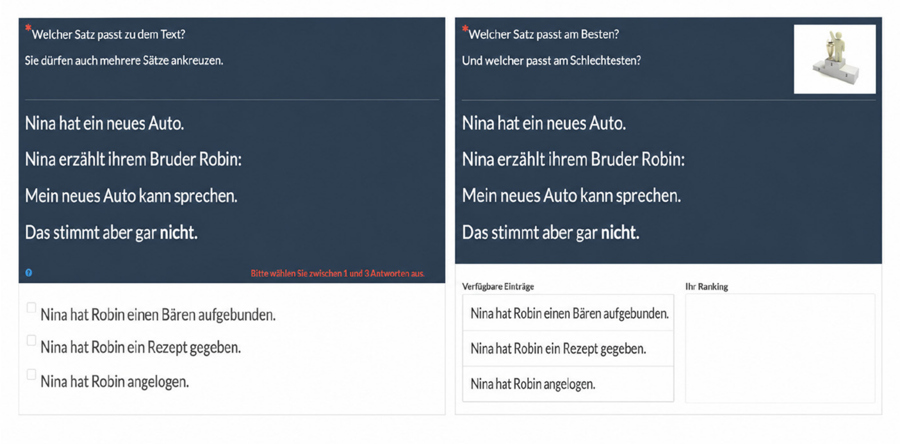


FIGURE 5
Example trial display for Experiment 4 (selection and ranking). The screenshots illustrate the two-step structure of the contextualized task. Following presentation of a short Easy German context (top), participants first completed a multiple-choice selection task by indicating which sentence or sentences matched the text (left). The three response options consisted of an idiom sentence, a meaning-consistent literal paraphrase, and an unrelated distractor. Participants then ranked the same three options using a podium metaphor (gold, silver, bronze; right), which was intended to make ordinal preference judgments more concrete for the target population.

### 6.2.3 Two-step response format (selection + ranking)

Each trial comprised two consecutive tasks. In Step 1 (multiple choice), participants were asked, *Welcher Satz, oder welche Sätze passen zum Text?* ("Which sentence, or which sentences, match the text?"). Multiple selections were permitted. This was intended to reflect a choose-all-that-apply format commonly used in accessibility-oriented testing. In Step 2 (ranking/preferences), participants ranked the same three options using a podium metaphor (gold/silver/bronze). The best-fitting option was assigned gold, the second-best silver, and the least-fitting bronze. We chose this sports-based metaphor to make the otherwise abstract notion of ordinal ranking more concrete and to reduce metalinguistic and numeracy demands. Participants indicated their ranking orally or by pointing, and the experimenter entered the choices.

### 6.2.4 Practice trials and comprehension scaffolding

Before the main task, participants completed two practice trials that matched the full structure of the experimental trials (context → selection → ranking). During these practice trials, participants were explicitly encouraged to ask the mediator as many questions as needed about the procedure and response format. The main task began only once participants indicated that they had no further questions. The podium metaphor was introduced and practiced separately during the context-comprehensibility check with members of the same target population, in order to ensure that the ranking procedure and the meaning of "best fit," "second-best fit," and "worst fit" were understood reliably.

### 6.2.5 Mediator-based responding

As in Experiment 3, responses were mediated by the experimenter to minimize barriers related to mouse use, device familiarity, and motor demands. The experimenter confirmed each recorded response aloud to ensure that the entered selection/ranking matched the participant's intended choice.

## 6.3 Dependent measures

Because Experiment 4 separates recognition/endorsement from preference, we distinguish two outcome families: (i) multiple-choice endorsement of the idiom option in context and (ii) ordinal preference rankings among the same three alternatives.

### 6.3.1 Multiple-choice endorsement outcomes

The selection task allowed multiple responses ("choose all that fit"). We therefore operationalized comprehension-related behavior at the level of endorsement rather than forcing a single-choice accuracy metric. The primary outcome for inferential analyses is: IdiomSelected (0/1); whether the participant endorsed the idiom option on a given trial. This measure directly targets the key question for Easy German practice: under contextual support, what makes participants choose the idiom as an acceptable continuation/summary of the vignette? As complementary (secondary) outcomes, endorsement can be coded analogously for the other option types: ParaphraseSelected (0/1) and DistractorSelected (0/1). However, in our dataset distractor endorsements were sparse, which creates quasi-separation and can make logistic mixed models unstable (see Section 6.5).

#### 6.3.1.1 Preference ranking outcome

After selection, participants ranked the three options using a podium metaphor. Rankings were treated as ordered responses (gold/silver/bronze). In the primary preference model, the key predictor is: OptionType (idiom, literal paraphrase, distractor), with idiom set as the reference level, so that model coefficients directly quantify whether paraphrases or distractors are preferred relative to idioms.

## 6.4 Statistical analysis

### 6.4.1 Multiple-choice: predictors of idiom endorsement

To identify which factors motivate idiom endorsement in context, we fitted generalized linear mixed-effects models (binomial family, logit link, Bates et al. 2015) to trial-level IdiomSelected (1/0). We began with a maximal, theory-motivated fixed-effects structure that included literacy (centered continuous score), functional familiarity from Experiment 3 (participant × item; 1/0), item transparency (centered mean), item plausibility (centered mean), and the theoretically relevant interactions (in particular, transparency × plausibility, and interactions of these item factors with literacy and familiarity). We included random intercepts for participants and items. Because the dataset is relatively small (26 participants × 24 items) and some outcomes were sparse (especially distractor endorsement), the maximal models did not always converge with more complex random-effects structures. We therefore applied backward model selection, starting from the maximal model and iteratively removing the least informative terms, while keeping the key theoretically motivated predictors (literacy, functional familiarity, and transparency/plausibility) in the model. Model reduction was guided by convergence diagnostics and likelihood-based model comparison. The final model reported for idiom endorsement retained centered literacy (`score_c`), functional familiarity (`accuracy_sum`), and centered transparency (`tran_mean_c`), with random intercepts for participants and items.

### 6.4.2 Ranking: preference structure

Ranking responses were collected immediately after the multiple-choice step using a podium metaphor (gold/silver/bronze), yielding an ordered outcome for each option (idiom, literal paraphrase, distractor). We analyzed these data with an ordinal model (Venables and Ripley, 2002) that preserved the ordered nature of the response.

### 6.4.3 Primary ordinal model

For the main ranking analysis, we fitted an ordinal logistic regression (cumulative link/proportional odds) with ResponseType as the key predictor (idiom, literal paraphrase, distractor; idiom as the reference level). This model tests whether the rank distribution differs by response type.

### 6.4.4 Maximal model and backward selection

To test whether ranking preferences were modulated by participant and item-level factors, we additionally fitted a maximal, theory-motivated ordinal model (Venables and Ripley, 2002) including: participant literacy (centered continuous score; `score_c`), functional familiarity from Experiment 3 at the participant × item level (binary; `accuracy_sum`), item-level transparency (centered mean; `tran_mean_c`), item-level literal plausibility (centered mean; `plaus_mean_c`) and the theoretically relevant interactions (e.g., transparency × plausibility, and interactions with literacy and/or familiarity) Because the dataset is modest (26 participants × 24 items), we used backward model selection: starting from the maximal fixed-effects structure and iteratively removing the least informative terms while retaining the core predictors of interest (literacy, functional familiarity, transparency/plausibility). Model reduction was guided by convergence/stability and likelihood-based comparisons. Dropping the remaining non-significant covariates did not change the size/significance of the familiarity effect.

## 6.5 Results, multiple-choice endorsement in context

### 6.5.1 Model fit and random effects

The model was fitted to 624 observations (26 participants × 24 items). Random intercept variance was larger for participants than items (ID: Var = 1.27, SD = 1.13; item: Var = 0.20, SD = 0.45), indicating substantial between-participant heterogeneity in overall idiom endorsement rates.

### 6.5.2 Familiarity

Familiarity significantly increased idiom endorsement ($\beta$ = 0.70, SE = 0.31, $z$ = 2.24, $p$ = 0.025; Table 3). This indicates that stored idiom knowledge supports meaning recognition in context, even in low-literacy readers.

### 6.5.3 Literacy and transparency

Literacy ($\beta$ = 0.50, SE = 0.27, $z$ = 1.86, $p$ = 0.063) and transparency ($\beta$ = 0.33, SE = 0.19, $z$ = 1.76, $p$ = 0.078) showed marginal positive effects in the expected direction but did not reach conventional significance. Figure 6 illustrates accuracy patterns by familiarity and transparency across three literacy subgroups used for visualization only. The attenuation of plausibility effects relative to Experiment 3 suggests that contextual scaffolding and externally provided alternatives reduce literal–figurative competition at the level of meaning recognition.

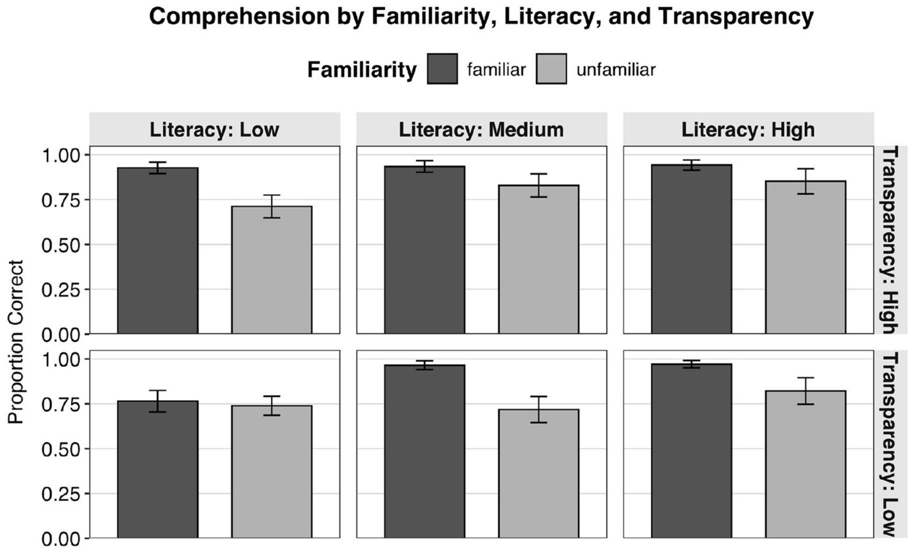


FIGURE 6
Comprehension accuracy in Experiment 4 as a function of familiarity and transparency, stratified by literacy (visualization only). Bars show the proportion of correct responses in the multiple-choice task (selection of both meaning-consistent options, i.e., the idiom and the literal paraphrase, with rejection of the distractor) for familiar vs. unfamiliar idioms. Panels on the *x*-axis display literacy subgroups (low, medium, high) created by splitting the continuous literacy score into three bins for visualization. Rows show transparency category (high vs. low). Error bars indicate uncertainty around the mean proportion. All inferential analyses reported in the text treat literacy as a continuous predictor.

### 6.5.4 Interpretation

The key result is that idiom endorsement in context is strongly shaped by whether the idiom was already accessible without context (functional familiarity from Experiment 3). The weak-to-marginal positive trends for literacy and transparency are compatible with the idea that both reader resources and item inferability may further support idiom endorsement, but the present data do not support strong claims about these factors in this task format.

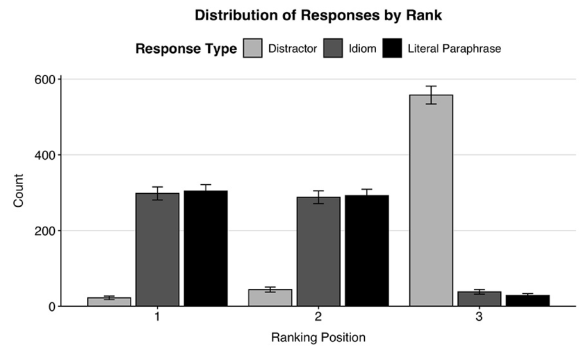


FIGURE 7
Distribution of ranking responses in Experiment 4 by option type. Bars show the number of times each response type (idiom, literal paraphrase, distractor) was assigned each ranking position (1 = best fit, 2 = second-best, 3 = worst fit) in the preference-ranking task following the contextualized multiple-choice trial. Error bars indicate variability across participants. The pattern shows robust rejection of distractors (predominantly ranked 3) and broadly comparable rankings for idioms and literal paraphrases, consistent with the ordinal regression results.

## 6.6 Results, ranking preferences

The ranking task provides a graded measure of how participants evaluated the three response options once they had been presented side by side in a supportive context. Ranking responses (gold/silver/bronze) were analyzed with an ordinal logistic regression (cumulative link model), with ResponseType entered as a categorical predictor and idioms as the reference level.

### 6.6.1 Option-type effects

The model shows a highly robust separation between context-inappropriate and meaning-consistent options. Distractor paraphrases were ranked substantially worse than idioms (ResponseType = Distractor: $\beta = 4.73$, SE = 0.18, $t = 26.44$, $p = 4.79 \times 10^{-154}$), indicating that participants were strongly sensitive to contextual fit and reliably rejected unrelated meanings. By contrast, literal paraphrases did not differ reliably from idioms (ResponseType = Literal: $\beta = -0.064$, SE = 0.11, $t = -0.59$, $p = 0.558$), suggesting that, when meaning was available, participants treated idiomatic and literal formulations as broadly equivalent in their preference rankings rather than systematically favoring one over the other. Because distractor ranks were highly skewed (Figure 7), we interpret the ordinal effect primarily as evidence for robust distractor rejection rather than as fine-grained sensitivity to rank differences among meaning-consistent options (Table 4).

TABLE 4 Ordinal regression coefficients for ranking responses (Experiment 4).

| Predictor | $\beta$ | SE | $t$ | $p$-value |
|---|---|---|---|---|
| ResponseType: distractor | 4.727 | 0.179 | 26.440 | <0.001 |
| ResponseType: literal paraphrase | −0.064 | 0.109 | −0.585 | 0.558 |
| Threshold 1–2 | −0.075 | 0.079 | −0.952 | 0.341 |
| Threshold 2–3 | 2.624 | 0.123 | 21.300 | <0.001 |

Idiom is the reference level for ResponseType.

### 6.6.2 Conditional preference analysis

To sharpen the preference question, we ran an additional analysis that treats preference as defined only for trials in which participants clearly rejected the distractor (i.e., assigned the distractor the lowest rank/bronze). This conditionalization reduces the decision to a direct competition between the idiom and the literal paraphrase, allowing us to ask: given that the distractor was treated as inappropriate, what predicts whether participants place the idiom above the literal paraphrase (vs. the reverse)? We created a binary outcome (RankBinary) with RankBinary = 1 when the idiom was ranked higher (better) than the literal paraphrase, and RankBinary = 0 otherwise. As stated above, we then fitted a mixed-effects logistic regression (binomial logit, Bates et al. 2015) with random intercepts for participants and items. The conditional dataset comprised 558 trials (out of 624), i.e., those trials where the distractor was assigned the lowest rank. Random-intercept variance was small but non-zero at both grouping levels (participants: Var = 0.083, SD = 0.289; items: Var = 0.082, SD = 0.286), indicating limited heterogeneity in baseline preference once the analysis was restricted to trials in which the distractor had been rejected. Crucially, functional familiarity from Experiment 3 (`accuracy_sum`) was the only reliable predictor: $\beta = 0.428$, SE = 0.200, $z = 2.146$, $p = 0.032$. Thus, when an idiom's meaning was already accessible without contextual support in Experiment 3, participants were more likely in Experiment 4 to rank the idiom above the literal paraphrase. Expressed as an odds ratio, this effect corresponds to $\exp(0.428) \sim 1.53$, that is, approximately 53% higher odds of preferring the idiom over the literal paraphrase when functional familiarity was present. By contrast, literacy (`score_c`) was not a significant predictor ($p = 0.599$), transparency showed only a non-significant negative trend ($p = 0.132$), and literal plausibility had an effect close to zero ($p = 0.679$). Taken together, these results suggest that, once comprehension is secured and the distractor is rejected, preferences between two meaning-consistent options are driven primarily by prior availability of the idiomatic meaning rather than by item-internal cue structure (Table 5).

TABLE 5 Mixed-effects logistic regression predicting idiom-over-literal preference (RankBinary) in Experiment 4 (subset with distractor ranked last).

| Predictor | $\beta$ | SE | $z$ | $p$-value |
|---|---|---|---|---|
| (Intercept) | −0.033 | 0.122 | −0.274 | 0.784 |
| Literacy (`score_c`) | −0.058 | 0.110 | −0.526 | 0.599 |
| Functional familiarity (`accuracy_sum`) | 0.428 | 0.200 | 2.146 | 0.032 |
| Transparency (`tran_mean_c`) | −0.171 | 0.114 | −1.505 | 0.132 |
| Literal plausibility (`plaus_mean_c`) | −0.035 | 0.085 | −0.413 | 0.679 |

## 6.7 Interim discussion

Experiment 4 supports two core conclusions. First, familiarity is a central facilitator of idiom endorsement in context. This aligns with the view that many idioms are stored and can be retrieved as multiword units once a supportive context triggers access. Second, idioms are not categorically avoided. When the idiom's meaning is available, participants treat idioms and paraphrases as similarly acceptable, while reliably rejecting distractors.

From an applied perspective, this suggests that paraphrases can function as scaffolding rather than as mandatory replacements. In other words, Easy German writers may be able to preserve idiomatic expressions that are familiar to the audience while providing paraphrastic support when needed. The lack of strong plausibility and transparency effects in comprehension and preference is also informative. Under contextual support and a recognition-based format, readers may rely more on stored knowledge and contextual fit than on analytic decomposition. This complements the findings of Experiment 3, where open-ended explanation exposed strong plausibility-driven competition.

# 7 General discussion

The project set out to evaluate whether adult low-literacy readers can process idioms in Easy German and to identify which factors predict comprehension and acceptability. Across four linked studies, we find that idiom accessibility is not a categorical property of figurative language. Instead, it reflects the interaction of reader skill, stored idiom knowledge, item-level properties that strengthen literal competition, and the availability of contextual and task-based scaffolding. Below we summarize the main findings in relation to the four hypotheses and then return to broader theoretical and practical implications.

## 7.1 Literacy effects

H1 predicted that higher literacy would support idiom comprehension. This was strongly supported in Experiment 3, where literacy predicted the probability of generating a correct figurative interpretation in the absence of context. The result is consistent with the idea that literacy captures not only decoding but also vocabulary breadth, integration skills, and executive resources that support the maintenance and evaluation of competing interpretations. In Experiment 4, literacy effects were weaker and only marginal. This pattern is theoretically meaningful rather than problematic. Contextual support and constrained response formats reduce the requirement to self-generate and articulate meaning, and thus reduce the extent to which literacy differences dominate performance. In other words, Easy German-style scaffolding appears to partially achieve its intended function of reducing ability-related variance. For future work, it would be valuable to further decompose literacy into component skills that may differentially relate to idiom processing, for example word-level decoding vs. sentence-level integration. The current study treats literacy as a unitary continuous index, which is appropriate for modeling but does not identify which subskills drive the effect.

## 7.2 Familiarity and retrieval

H2 predicted that familiarity would facilitate comprehension, particularly under contextualized demands. Experiment 3 established that baseline access varies widely across idioms and participants, confirming the practical point that idiom familiarity cannot be assumed in the target group. Experiment 4 then showed a reliable positive effect of familiarity on comprehension accuracy. This pattern supports a retrieval-based contribution to idiom processing. When an idiom is part of a reader's stored repertoire, context can trigger rapid access to figurative meaning. Importantly, the acceptance of idioms in the ranking task suggests that retrieval does not merely enable comprehension, it also licenses the idiom as an acceptable communicative form. From the standpoint of Easy German practice, this implies that audience familiarity should be treated as an empirical variable. A transparent idiom that is unfamiliar may still fail, while an opaque but highly familiar idiom may succeed. Consequently, guideline statements that focus on transparency without addressing familiarity risk being unreliable.

## 7.3 Literal plausibility and competition

H4 predicted that high literal plausibility would strengthen literal interpretations and reduce figurative accuracy. This was strongly supported in Experiment 3. Plausibility increased literal responses and reduced figurative interpretations when idioms were presented without context. This plausibility effect can be interpreted as evidence for sustained competition rather than for a simple serial 'literal-first then figurative' process. The key point is that plausibility does not merely delay the recognition of figurativity, it provides a coherent endpoint for interpretation. In no-context settings, a plausible literal interpretation can be good enough, especially when the task does not provide incentives to search for an alternative. In Experiment 4, plausibility effects were attenuated. This suggests that context and response options can dampen competition, either by making the figurative candidate more salient or by making literal interpretation less relevant to the discourse. In applied terms, this highlights the importance of contextual licensing when idioms might otherwise invite plausible literal readings.

## 7.4 Transparency as a conditional cue

H3 predicted that transparent idioms would be easier to interpret. The results do not support a simple main-effect version of this hypothesis. In Experiment 3, transparency did not facilitate comprehension and instead interacted with plausibility, with the transparent and plausible condition showing the lowest figurative accuracy. One way to interpret this pattern is that transparency has two faces. It can support figurative inference by providing a bridge from literal components to figurative meaning, but it can also strengthen literal composition by making the literal scene coherent. When literacy and task demands limit analytic mapping, the literal-supporting face may dominate. Experiment 4 shows that transparency may have weak positive effects under contextual support, but these were only marginal. Together, the results suggest that transparency is not a reliable stand-alone accessibility criterion for adult low-literacy readers. Instead, its effect depends on the strength of literal competition and the availability of contextual and task-based scaffolding. This conditional view offers a concrete reinterpretation of guideline discourse. Rather than advising "use transparent idioms," guidance might instead consider whether transparency is paired with high literal plausibility and whether context makes the figurative meaning explicit enough to prevent a literal settling point.

## 7.5 Task format and methodological implications

The contrast between Experiments 3 and 4 underscores a general methodological point. Task format is not a neutral window onto competence. Open-ended explanation tasks place high demands on abstraction, planning, and expressive language. Multiple-choice tasks reduce expressive demands and provide external scaffolding. In our data, the no-context explanation task exposed strong literacy and plausibility effects and revealed that transparency is not automatically usable. The contextualized multiple-choice task showed strong familiarity effects and robust distractor rejection, indicating that participants can make fine-grained contextual fit judgments when the task supports them. This has practical implications for accessibility research. When testing target groups, the experiment must itself be accessible. Otherwise, task demands can masquerade as comprehension deficits. The present study's mediated response procedure, large-font design, and ranking metaphor are part of the methodological contribution of the work.

## 7.6 Theoretical positioning

Although our hypotheses were formulated at the level of cue effects, the results bear on broader processing accounts. The strong familiarity effect in context is consistent with retrieval-based and hybrid models that posit lexicalized representations of idioms (Swinney and Cutler, 1979; Glucksberg, 2001; Titone and Libben, 2014). The strong plausibility effect in no-context explanation is consistent with constraint-based competition accounts (Titone and Connine, 1999; Libben and Titone, 2008; Titone and Libben, 2014) in which literal and figurative candidates remain active and compete for selection. The transparency by plausibility interaction suggests that compositional coherence can strengthen the literal competitor, which is naturally captured in constraint-based accounts but is less straightforward for accounts that treat transparency as a uniformly facilitative route to figurative meaning. At the same time, the present data do not adjudicate fine-grained architectural claims about time course because the tasks measure end-state outcomes rather than online processing. Future work could extend this research to self-paced reading or eye-tracking in Easy German contexts, with accessibility-adapted procedures, to examine when competition emerges and how it is resolved.

### 7.7 Limitations and future directions

Several limitations are inherent to working with protected institutional settings. Sample sizes are modest, attrition across sessions is nontrivial, and participant profiles are heterogeneous. Second, functional familiarity in Experiment 3 is established through participants' explanations, which introduces expressive demands. Some participants may know an idiom but struggle to verbalize its meaning or provide a fitting situation the idiom might occur in. Experiment 4 partially addresses this by providing recognition-based options. Third, our item set is necessarily limited. A larger item set would allow random effects for items and more nuanced modeling of continuous transparency and plausibility. However, expanding item sets must be balanced against participant fatigue. Finally, the literal construals observed in Experiment 3 are reminiscent of *concretism* — a tendency toward literal interpretation of figurative language that has been described, in the clinical literature on schizophrenia, as a bias toward the concrete aspects of stimuli (Bambini et al., 2025) — though we do not treat the target group as clinical and imply no clinical parallel. The goal is not diagnosis but understanding cue use under literacy constraints. Future studies could compare low-literacy readers to matched controls to more precisely separate literacy effects from general population variability.

## 8 Practical conclusions for Easy German

The results motivate a shift from categorical prohibitions toward conditional, evidence-based guidance. First, prioritize familiarity. Idioms that are functionally familiar to the audience can be comprehended and are not dispreferred relative to paraphrases. Where possible, familiarity should be assessed empirically, either through community norming or through domain-specific corpora of Easy German usage. Second, avoid high literal plausibility in low-context settings. Idioms whose literal scenes are easy to imagine pose a heightened risk of literal interpretation, especially when context is minimal. If such idioms are used, contextual cues should clearly license the figurative meaning and reduce the attractiveness of the literal candidate. Third, treat transparency with caution. Transparency is not a guarantee of accessibility. In the present data, transparent and plausible idioms were particularly difficult without context. Guidance that endorses transparent idioms should therefore be qualified, for example by considering plausibility and by ensuring contextual support. Fourth, use paraphrases as scaffolding rather than replacement. The ranking results suggest that readers accept idioms when meaning is available, and that paraphrases remain accessible. In practice, this motivates a strategy of pairing an idiom with a short paraphrase when the idiom serves a discourse function that a purely literal version might lose. Finally, incorporate testing into guideline development. Easy German rule sets are most useful when they are empirically grounded. The present study illustrates one approach to testing, combining norming, literacy assessment, baseline access measures, and contextual preference tasks.

## 9 Conclusion

This study began from an applied problem, Easy German guidelines provide ambivalent advice about figurative language, yet there is little evidence on how low-literacy adults process idioms. By combining norming, literacy assessment, a no-context diagnostic of baseline idiom access, and a contextualized comprehension and preference paradigm, we provide a detailed picture of the factors that shape idiom accessibility. Across experiments, three findings stand out. Literacy reliably supports figurative success when tasks require self-generated meaning. Familiarity is a robust facilitator of comprehension in context and licenses idioms as acceptable choices. Literal plausibility is the key competitor in no-context settings, and it interacts with transparency in ways that challenge the intuition that transparent idioms are necessarily safer. Taken together, the results support a cue-based competition view in which literal and figurative interpretations compete and in which the usability of cues depends on reader skills and task demands. For Easy German practice, the data argue against treating transparency as the primary criterion for idiom inclusion. Instead, accessibility is better predicted by whether an idiom is familiar to the audience and by whether the literal scene is sufficiently implausible or contextually disfavored. With contextual scaffolding and strategic paraphrasing, idioms can be both comprehensible and acceptable, suggesting that figurative language need not be categorically excluded from accessible communication.

## Data availability statement

The original contributions presented in the study are publicly available. The datasets and R scripts used for the statistical analyses reported in this article are available on the Open Science Framework: https://osf.io/83sdb.

## Ethics statement

The studies involving humans were approved by the Ethics Committee of the Linguistic Society of Germany (DGfS) on 22 April 2022 covered by ethics vote #2022-03-220422. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

## Author contributions

LW: Writing – original draft, Formal analysis, Visualization, Investigation, Data curation, Writing – review & editing. IR: Supervision, Methodology, Conceptualization, Resources, Funding acquisition, Writing – review & editing, Project administration.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. This study was funded by Gefördert durch die Deutsche Forschungsgemeinschaft (DFG)-Projektnummer 232722074-SFB 1102 / Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)-Project-ID 232722074-CRC 1102.

## Acknowledgments

We thank Anna Riga for her valuable assistance in preparing and conducting the experiments, and for her dedication and support at every stage of the project. We are also grateful to Robin Lemke for his thoughtful feedback and many helpful discussions; his constructive comments and continued encouragement throughout the process were greatly appreciated.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Bambini, V., Frau, F., Bischetti, L., Agostoni, G., Mevio, C., Battaglini, C., et al. (2025). From semantic concreteness to concretism in schizophrenia: an automated linguistic analysis of speech produced in figurative language interpretation. *Clin. Linguist. Phon.* 39, 1070–1092. doi: 10.1080/02699206.2025.2451961

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bock, B. M. (2019). "*Leichte Sprache*"*-Kein Regelwerk: Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt, volume 5 of Kommunikation-Partizipation-Inklusion*. Berlin: Frank and Timme. German.

Bock, B. M., and Pappert, S. (2023). *Leichte sprache, einfache sprache, verständliche sprache*. Tübingen: Narr Francke Attempto. German. doi: 10.24053/9783823391814

Bonin, P., Méot, A., and Bugaiska, A. (2013). Norms and comprehension times for 305 french idiomatic expressions. *Behav. Res. Methods* 45, 1259–1271. doi: 10.3758/s13428-013-0331-4

Borghardt, L., Deilen, S., Fuchs, J., Gros, A.-K., Hansen-Schirra, S., Nagels, A., et al. (2021). Neuroscientific research on the processing of easy language. *Front Commun.* 6:698044. doi: 10.3389/fcomm.2021.698044

Bredel, U., and Maaß, C. (2016). *Leichte sprache: theoretische grundlagen, orientierung für die praxis*. Berlin: Dudenverlag. German.

Cacciari, C., and Glucksberg, S. (1991). "Understanding idiomatic expressions: the contribution of word meanings," in *Understanding Word and Sentence*, eds G. B. Simpson (Amsterdam; North-Holland: Elsevier Science Publishers), 77, 217–240. doi: 10.1016/S0166-4115(08)61535-6

Cacciari, C., and Tabossi, P. (eds.) (1988). *Idioms: Processing, Structure, and Interpretation*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cain, K., Towse, A. S., and Knight, R. S. (2005). The relation between children's reading comprehension level and their comprehension of idioms. *J. Exp. Child Psychol.* 90, 65–87. doi: 10.1016/j.jecp.2004. 09.003

Cap, F., Ehren, R., Köper, M., Lichte, T., Schulte im Walde, S., and Zinsmeister, H. (2018). "VerbCompoCor: A German Corpus With Compositionality Judgments For Verb-Dependent Pairs," in *40th Annual Meeting of the DGfS, Universität Stuttgart* (Stuttgart).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cronk, B. C., and Schweigert, W. A. (1992). The comprehension of idioms: the effects of familiarity, literalness, and usage. *Appl. Psycholinguist* 13, 131–146. doi: 10.1017/S0142716400005531

DIN (2025). *DIN SPEC 33429: Leichte Sprache-Anforderungen an die Erstellung von Texten in Leichter Sprache. Technical report*. Berlin: Deutsches Institut für Normung. German.

Donalies, E. (2012). *Basiswissen Deutsche Phraseologie*. Tübingen: Narr. German.

Dudenredaktion (2021). *Duden-Redewendungen: Wörterbuch der deutschen Idiomatik, 5th Edn*. Berlin: Dudenverlag. German.

Ehren, R., Lichte, T., Kallmeyer, L., and Waszczuk, J. (2020). "Supervised disambiguation of German verbal idioms with a BiLSTM architecture," in *Proceedings of the second workshop on figurative language processing*

(Association for Computational Linguistics), 211–220. doi: 10.18653/v1/2020.figlang-1.29

Gibbs, R. W., Nayak, N. P., and Cutting, C. (1989). How to kick the bucket and not decompose: analyzability and idiom processing. *J. Mem. Lang.* 28, 576–593. doi: 10.1016/0749-596X(89)90014-4

Glucksberg, S. (2001). *Understanding Figurative Language: From Metaphors to Idioms*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195111095.001.0001

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics, Vol. 3: Speech Acts*, eds P. Cole, and J. L. Morgan (New York, NY: Academic Press), 41–58. doi: 10.1163/9789004368811_003

Koppel, I., and Wolf, K. D. (2014). Otu.lea: Eine niedrigschwellige Online-Diagnostik für funktionale Analphabet(inn)en in der Kursarbeit. *Alfa-Forum*, 86, 28–31. German.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Levorato, M. C., Nesi, B., and Cacciari, C. (2004). Reading comprehension and understanding idiomatic expressions: a developmental study. *Brain Lang.* 91, 303–314. doi: 10.1016/j.bandl.2004.04.002

Libben, M. R., and Titone, D. A. (2008). The multidetermined nature of idiom processing. *Mem. Cogn.* 36, 1103–1121. doi: 10.3758/MC.36.6.1103

Morid, M., and Sabourin, L. (2024). Affective and sensory-motor norms for idioms by L1 and L2 English speakers. *Appl. Psycholinguist* 45, 138–155. doi: 10.1017/S0142716423000504

Nation, K. (2005). "Children's reading comprehension difficulties," in *The Science of Reading: A Handbook*, eds M. J. Snowling, and C. Hulme (Oxford: Blackwell), 248–266. doi: 10.1111/b.9781405114882.2005.00019.x

Netzwerk Leichte Sprache (2022). *Die Regeln für Leichte Sprache*. German. Available online at: https://www.netzwerk-leichte-sprache.de/fileadmin/content/documents/regeln/Regelwerk_NLS_Neuauflage-2022.pdf Accessed June 16, 2026.

Nippold, M. A. (1991). Evaluating and enhancing idiom comprehension in language-disordered students. *Lang. Speech Hear. Serv. Sch.* 22, 100–106. doi: 10.1044/0161-1461.2203.100

Nippold, M. A. (1998). *Later Language Development: The School-Age and Adolescent Years, 2nd Edn*. Austin, TX: Pro-Ed.

Nippold, M. A., and Rudzinski, M. (1993). Familiarity and transparency in idiom explanation: a developmental study of children and adolescents. *J. Speech Hear. Res.* 36, 728–737. doi: 10.1044/jshr.3604.728

Nippold, M. A., and Taylor, C. L. (2002). Judgments of idiom familiarity and transparency: a comparison of children and adolescents. *J. Speech Lang. Hear. Res.* 45, 384–391. doi: 10.1044/1092-4388(2002/030)

Nordmann, E., Cleland, A. A., and Bull, R. (2014). Familiarity breeds dissent: reliability analyses for British-English idioms on measures of familiarity, meaning, literality, and decomposability. *Acta Psychol.* 149, 87–95. doi: 10.1016/j.actpsy.2014.03.009

Palan, S., and Schitter, C. (2018). Prolific.ac – a subject pool for online experiments. *J. Behav. Exp. Finance* 17, 22–27. doi: 10.1016/j.jbef.2017.12.004

Perfetti, C. A. (1985). *Reading Ability*. New York, NY: Oxford University Press.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rink, I., Maaß, C., Hansen-Schirra, S., and Gutermuth, S. (2023). "Proband(inn)en mit besonderen kommunikativen bedarfen in der empirischen datenerhebung," in *Gesundheitskompetenz*, eds K. Rathmann, K. Dadaczynski, O. Okan, and M. Messer (Berlin/Heidelberg: Springer), 97–108. German. doi: 10.1007/978-3-662-67055-2_147

Searle, J. R. (1979). "Metaphor," in *Metaphor and Thought*, ed A., Ortony (Cambridge: Cambridge University Press), 92–123. doi: 10.1017/CBO9780511609213.006

Swinney, D. A., and Cutler, A. (1979). The access and processing of idiomatic expressions. *J. Verbal Learn. Verbal Behav.* 18, 523–534. doi: 10.1016/S0022-5371(79)90284-6

Tabossi, P., Arduino, L. S., and Fanari, R. (2011). Descriptive norms for 245 Italian idiomatic expressions. *Behav. Res. Methods* 43, 110–123. doi: 10.3758/s13428-010-0018-z

Titone, D. A., and Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: familiarity, compositionality, predictability, and literality. *Metaphor Symb. Act.* 9, 247–270. doi: 10.1207/s15327868ms0904_1

Titone, D. A., and Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *J. Pragmat.* 31, 1655–1674. doi: 10.1016/S0378-2166(99)00008-9

Titone, D. A., and Libben, M. R. (2014). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom meaning activation: a cross-modal priming investigation. *Ment. Lex.* 9, 473–496. doi: 10.1075/ml.9.3.05tit

Vega-Moreno, R. E. (2007). Idioms, transparency and pragmatic inference. *UCL Work. Pap. Linguist.* 19, 373–398.

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics With S, 4th Edn*. New York, NY: Springer. doi: 10.1007/978-0-387-21706-2

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511519772

Zehr, J., and Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. doi: 10.17605/OSF.IO/MD832