Which Words Are Most Iconic, and Which Are Rated Most Consistently? A Large-Scale Analysis of Human Iconicity Ratings with Imputed Predictors

Hinano Iida (iida.hinano.i1@s.mail.nagoya-u.ac.jp)[1] & Ryo Hori[2]
[1]Nagoya University
[2]Sugiyama Jogakuen University

Abstract

Iconicity is a central notion in cognitive science, attracting growing attention from perspectives on language acquisition, processing, and language evolution. While previous research has emphasized the subjective nature of iconicity, quantitative analyses of iconicity ratings have relied primarily on mean values, treating iconicity as a stable lexical property, and have often been constrained by substantial data loss when integrating multiple lexical and psycholinguistic datasets. The present study addresses these limitations by analyzing both mean iconicity ratings and inter-rater variability across 14,764 English words, combining a conservative complete-case analysis with complementary analyses based on imputed predictors to increase lexical coverage. For example, in the complete-case analysis, earlier-acquired and etymologically imitative words tended to receive higher mean iconicity ratings, whereas rating variability was greater for earlier-acquired and more familiar words. By examining both mean ratings and variability, we distinguish factors associated with perceived iconicity from those associated with convergence or divergence in iconicity judgments, offering a broader account of lexical iconicity and heterogeneity in iconic perception.

Keywords: Iconicity; Iconicity ratings; individual differences; data imputation; machine-learning–based imputation

Introduction

Iconicity—the resemblance between linguistic form and meaning—is increasingly recognized as a foundational property of human language across modalities, including both spoken and signed languages (Dingemanse et al., 2015; Ferrara & Hodge, 2018; Perniss et al., 2010), and has attracted growing attention from multiple perspectives, including word learning (Imai & Kita, 2014; Ortega, 2017; Ortega et al., 2017; Perry et al., 2015, 2018, 2021; Thompson et al., 2012; Vinson et al., 2008), language processing (Bosworth & Emmorey, 2010; Perniss & Vigliocco, 2014; Sidhu et al., 2020; Thompson, 2011; Thompson et al., 2010), and language evolution (Ćwiek et al., 2021; Imai & Kita, 2014; Fay et al., 2014; Macuch Silva et al., 2020; Perlman et al., 2015; Perlman & Lupyan, 2018).

A recent definition characterizes iconicity as a property of signals in any modality or medium whose production and/or interpretation involves *a sense of* resemblance between at least some aspect of form and meaning (Winter et al., 2026). Despite this emphasis on subjectivity and the growing body of cross-linguistic work on iconicity (e.g., Occhino et al., 2017), most large-scale studies of lexical iconicity within individual languages have focused on average iconicity ratings, effectively treating iconicity as a stable property of individual words (e.g., Perry et al., 2015; Winter et al., 2024).

As a result, inter-rater variability and the conditions under which iconicity judgments converge or diverge within a language community remain largely unexplored. This variability is theoretically important because it provides a way to ask when iconicity functions as a relatively shared cue and when subjective interpretation plays a larger role. A word with high mean iconicity may be judged as iconic for broadly shared reasons, or it may receive high ratings because some speakers perceive strong form–meaning resemblance whereas others do not. This distinction may have consequences for language development and processing: consistently perceived iconic mappings may provide more reliable cues for form–meaning mapping, whereas highly variable iconicity may depend more on individual experience, contextual support, or community-specific conventions. At the same time, previous large-scale analyses have also been constrained by a restricted set of predictors or substantial data loss when integrating multiple lexical and psycholinguistic datasets.

To address these limitations, the present study analyzes both average iconicity ratings and inter-rater variability across 14,764[1] English words. We combine a conservative complete-case analysis of fully observed predictors with complementary analyses based on imputed predictors to increase lexical coverage. This approach allows us to identify predictors of average iconicity as well as predictors of rating variability, thereby distinguishing factors associated with perceived iconicity from those associated with convergence or divergence in iconicity judgments.

Previous Work

Iconicity and Iconicity Ratings

Research on iconicity is characterized by substantial methodological diversity, spanning experimental paradigms, corpus-based approaches, and behavioral norming studies

[1] The original dataset contained 14,776 items, but Lily and lily were excluded because, although they were included in the iconicity ratings, it was unclear whether they were treated as proper nouns or common nouns in the other datasets. In addition, ten words categorized as "unclassified" in the SUBTLEX-US corpus (Brysbaert & New, 2009) were excluded.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

(Motamedi et al., 2019). Within this literature, iconicity has often been operationalized, among other approaches, through iconicity ratings, in which language users are asked to judge how much a linguistic form resembles its meaning. Because this method relies directly on speakers' intuitions, iconicity ratings are particularly well suited to capturing the subjective dimension of iconicity—that is, how strongly language users perceive a resemblance between form and meaning.

Iconicity rating experiments started in research on sign language: ASL (Bosworth & Emmorey, 2010; Caselli et al., 2017; Sehyr & Emmorey, 2019), ASL and German Sign Language (DGS) (Occhino et al., 2017), British Sign Language (BSL) (Vinson et al., 2008), and BSL and DGS (Ortega et al., 2025). They have become increasingly common in spoken language research as well: English and Spanish (Perry et al., 2015; Winter et al., 2017, 2024), Spanish (Hinojosa et al., 2021), Japanese (McLean et al., 2023; Thompson et al., 2020), Chinese (Van Hoey et al., 2024), and Italian (de Varda et al., 2025). These datasets have enabled quantitative investigations of lexical iconicity and have become a central resource for research on iconicity.

Relation Between Iconicity and Other Variables

Previous research has shown that iconicity ratings are systematically related to a wide range of psycholinguistic variables. Across English and signed languages such as ASL and BSL, iconicity tends to be higher for words and signs that are learned earlier in life (Caselli & Pyers, 2017; Massaro & Perlman, 2017; Perry et al., 2015, 2018; Thompson et al., 2012; Vinson et al., 2008; Winter et al., 2024). Iconicity ratings in English, Spanish, and ASL are also positively correlated with sensory experience ratings, indicating that words with stronger perceptual associations tend to be judged as more iconic (Hinojosa et al., 2021; Perlman et al., 2018; Sidhu & Pexman, 2018; Winter et al., 2017). In addition, English iconicity ratings correlate with funniness ratings and with measures of structural markedness, suggesting that iconic words often stand out from more typical lexical forms (Dingemanse & Thompson, 2020). Iconicity ratings are negatively correlated with word frequency when frequency is derived from adult language corpora (Perry et al., 2015; Winter et al., 2024). Furthermore, iconicity ratings differ systematically across parts of speech in English, Spanish, and Japanese (Hinojosa et al., 2021; Perry et al., 2015, 2018; Winter et al., 2024). With respect to concreteness, previous findings are mixed: while some studies suggest that more concrete concepts are more likely to be expressed iconically (Lupyan & Winter, 2018), others report negative correlations between them (Winter et al., 2024).

Although previous studies have identified several correlates of iconicity, existing large-scale analyses remain limited in two key respects. First, they have focused primarily on mean iconicity ratings, leaving open when iconicity judgments are shared across speakers and when they depend more strongly on subjective construal. Second, they have examined a relatively restricted set of lexical and psycholinguistic predictors, partly because integrating multiple norming datasets leads to substantial data loss. Building on Winter et al. (2024), the present study examines both mean iconicity and inter-rater variability across 14,764 English words, asking which factors are associated with perceived iconicity and which are associated with consistency or divergence in those judgments.

Method

Dataset

The present study integrates a large-scale English iconicity dataset with a broad range of lexical and psycholinguistic predictors to examine both the mean and the variability of iconicity ratings.

Iconicity ratings were taken from the large-scale English norming study reported by Winter et al. (2024). The dataset comprises ratings for 14,776[1] English words provided by 1,419 speakers of American English. Participants rated each word on a seven-point Likert scale ranging from 1 ("Not iconic at all") to 7 ("Very iconic"). The word list includes a wide range of lexical categories, including interjections, verbs, nouns, adjectives, adverbs, and function words.

Concreteness ratings were taken from Brysbaert et al. (2014), who collected ratings for 37,058 English words and 2,896 two-word expressions using a five-point scale ranging from abstract to concrete. *Imageability, familiarity, semantic size,* and *gender association* were taken from the Glasgow Norms, which provide ratings for 5,553 English words across nine dimensions (Scott et al., 2019). For dimensions for which larger-scale datasets were available (e.g., concreteness), ratings from alternative sources were used. *Valence* (pleasantness), *arousal* (emotional intensity), and *dominance* (degree of control) were taken from Warriner et al. (2013), who collected ratings for 13,915 English lemmas using nine-point scales.

Sensorimotor strength ratings for 39,707 concepts across six perceptual modalities and five action effectors were obtained from the Lancaster sensorimotor norms (Lynott et al., 2020). Ratings range from 0 (not experienced at all) to 5 (experienced greatly). We focused on the six *perceptual modality strength*, *perceptual exclusivity*, and *maximum perceptual strength*, which have been shown to be closely related to iconicity.

Cognition ratings, which index the degree to which a word is associated with mental activities or cognitive products (e.g., *reasoning*, *belief*), were obtained from a dataset of 8,826 English words rated on a seven-point Likert scale (Corenblum & Pexman, 2025). *Funniness* ratings were taken from Engelthaler and Hills (2018), who collected ratings for 4,997 English words on a five-point scale ranging from 1 (humorless) to 5 (humorous). *Age of acquisition* (AoA) ratings were taken from Kuperman et al. (2012). Participants reported the age (in years) at which they believed they had learned each word, defined as the age at which they would have understood the word if it had been used in context. The dataset includes 30,121 English content words (nouns, verbs, and adjectives).

Structural markedness was indexed using *log letter frequency*, following Dingemanse and Thompson (2020). *Word frequency* was measured using log-transformed SUBTLEX-US frequencies derived from a 51-million-word corpus of American English subtitles (Brysbaert & New, 2009). *Word length* was operationalized as character count, a factor known to influence word recognition (New et al., 2006). *Part-of-speech* (POS) information was taken from the SUBTLEX-US corpus (Brysbaert et al., 2012). For each word, the *dominant part of speech* was used. *Semantic neighborhood density* was measured using average radius of co-occurrence (ARC; Shaoul & Westbury, 2010), defined as the mean distance between a target word and its semantic neighbors within a specified threshold.

Phonological surprisal was operationalized as average phonemic bigram surprisal, computed using a large corpus of spoken American English (Kilpatrick & Bundgaard-Nielsen, 2025). Surprisal values are available for 42,223 words. *Etymological imitativeness* was coded based on a dataset developed by Iida and Akita (2023), which extracted all words labeled as onomatopoeic, echoic, or imitative in the etymology sections of the Oxford English Dictionary. Words whose etymologies were judged not to be genuinely imitative (e.g., *sand laverock*, spout, *Strine*) were manually excluded. Remaining words were coded as etymologically imitative.

When restricting analyses to words with complete data across all predictors, the resulting dataset comprised 1,071 words. To avoid substantial data loss and retain the full lexical coverage of the iconicity ratings, we additionally conducted analyses using imputed predictor values, allowing all 14,764 words from Winter et al. (2024) to be included. Details of the imputation procedure are described in the following section.

Imputation Procedure

To make full use of the currently available data, missing values are imputed using machine learning techniques. For imputation, we employ AutoGluon (version 1.3.1), an open-source software framework provided by Amazon (Erickson et al., 2020). AutoGluon is an AutoML (automated machine learning) system that enables the automated execution of machine learning workflows. By specifying a data frame and the target column to be predicted, AutoML automatically performs many of the processes required for machine learning, including feature engineering, model selection, and ensemble methods such as bagging, etc. This automation allows for the efficient development of high-performing machine learning models with minimal manual intervention[2].

The complete set of acquired datasets is referred to as the *original data*. For each imputation task, the column to be imputed is defined as the *target column*, and all remaining columns in the original data (excluding the target column) are defined as *feature columns*. The target column is predicted using the feature columns. Only values from the original data are used in the feature columns; imputed values are not used.

To evaluate imputation accuracy, 500 rows in which the target column contains human ratings are reserved as a test set. The remaining rows with target human ratings are used as training and validation data to construct several predictive models. The model that achieves the minimum validation loss is selected as the best model and is subsequently evaluated on the held-out test data. By comparing the training losses on the validation and test sets, we verify that the selected model does not suffer from overfitting. Finally, to obtain a higher-performance model trained on a larger amount of data, a final model is constructed using all rows in which the target column contains human ratings. This final model is then used to impute missing values in rows where the target column does not have human ratings. This procedure is repeated for all target columns subject to imputation.

As evaluation metrics, we use log loss for discrete target variables and mean absolute error (MAE) for continuous target variables. Both metrics have a theoretical minimum of 0 and an unbounded upper limit. The model search time of the AutoML is set to 600 seconds per target column[3]. The number of bagging folds (*num_bag_folds*) is set to 10, the number of stacking levels (*num_stack_levels*) is set to 2, and the *best_quality* preset is used to prioritize predictive performance. Bagging is an ensemble method that improves performance by constructing multiple models and aggregating their predictions via majority voting or averaging; in this study, ten models are trained simultaneously and aggregated. Stacking is a method that combines multiple heterogeneous models in a hierarchical manner, and we employ a two-level stacking configuration.

Statistical Analyses

Data preprocessing and imputation were conducted in Python, and all statistical analyses were performed in R[4] (version 4.4.0; R Core Team, 2024). To address our research questions, we fitted linear regression models predicting average iconicity and inter-rater variability from a broad set of lexical, psycholinguistic, affective, and structural predictors. Because rating variability on a bounded scale is partly constrained by an item's mean rating, the variability models additionally controlled for each word's average distance from the overall mean iconicity rating (Winter et al., 2024). All predictors were entered simultaneously, continuous predictors were standardized, and models were fitted to both raw and imputed datasets using identical specifications. Model assumptions, including the residual distributions of the SD models, were assessed by visual inspection of residual histograms and Q-Q plots, which did not indicate severe deviations from normality.

[2] Imputation environment is Python 3.12.10 on a MacBook Pro (CPU: M4 Pro, RAM: 48GB).

[3] We additionally tested 3,600 seconds but found that 600 seconds was sufficient.

[4] All code, data, and supplementary materials, including full regression outputs, are available at: https://osf.io/jhcsw.

Results

Table 1 provides a descriptive overview of words with the highest and lowest mean iconicity, as well as the highest and lowest inter-rater variability, in the raw and imputed datasets. Due to space limitations, selected imputation results are reported in Table 2. For all 55 imputed columns, the selected models were WeightedEnsembleL2, L3, or L4, which are final meta-models constructed by taking a weighted average of predictions from multiple machine learning models. To provide a reference for the magnitude of the training losses, the minimum and maximum values observed in the original data are also reported.

Table 1: Descriptive overview of words with extreme mean iconicity and rating variability.

	Raw	Imputed
Highest mean	whiff, shriek, crunch	oomph, swish, wiggle
Lowest mean	shape, request, chauffeur	how, if, partial
Highest variability	bread, town, football	smokehouse, effortless, smatter
Lowest variability	whiff, gloom, penguin	oomph, swish, wiggle

Mean Iconicity

Overall results for mean iconicity are reported in Table 3. The complete-case model accounted for 21% of the variance in mean iconicity ratings (adjusted R^2 = .21), and the full-coverage imputed model accounted for 28% of the variance (adjusted R^2 = .28). Adjusted GVIF values indicated modest but not severe multicollinearity, with maximum values of 3.13 in the complete-case models and 3.19 in the imputed models[5].

Table 2: Partial Imputation Results

Variable	Min	Max	MAE
Concreteness	1.04	5.00	0.22
Familiarity	1.65	6.90	0.27
Imageability	1.88	6.90	0.30
Semantic size	1.38	6.91	0.44
Valence	1.26	8.53	0.54
Arousal	1.60	7.79	0.53
Dominance	1.68	7.90	0.44
Auditory strength	0.00	5.00	0.07
Visual strength	0.00	5.00	0.04
Haptic strength	0.00	4.94	0.09
Olfactory strength	0.00	2.42	0.07
Interoceptive strength	0.00	4.90	0.07
Perceptual exclusivity	0.04	1.00	0.00
Max perceptual strength	0.56	5.00	0.01
Cognition rating	1.00	6.70	0.35
Funniness	1.18	4.32	0.22
Age of acquisition	1.58	19.21	1.10
Log letter frequency	–4.37	–2.20	0.13
Semantic neighborhood density	0.13	0.74	0.06
Phonemic surprisal	0.44	12.67	0.63

Table 3: Regression coefficients predicting iconicity ratings from all predictors. For the categorical predictors, the reference level for etymology was imitative origin, and the reference level for part of speech (POS) was noun. POS: Interjection was not estimated in the complete-case model because no interjections remained after complete-case filtering.

Predictors	Estimate (Raw)	*p* (Raw)		Estimate (Imputed)	*p* (Imputed)	
Valence	0.03	0.486		–0.09	< .001	***
Arousal	0.09	0.001	**	0.10	< .001	***
Dominance	–0.07	0.033	*	0.05	< .001	***
Log letter frequency	–0.08	< .001	***	–0.09	< .001	***
Concreteness	0.08	0.261		–0.04	0.064	.
Semantic neighborhood density	–0.19	< .001	***	–0.04	< .001	***
Funniness	0.08	0.013	*	0.11	< .001	***
Word frequency	–0.12	0.02	*	–0.12	< .001	***
Age of acquisition	–0.18	< .001	***	–0.23	< .001	***
Imageability	–0.06	0.313		0.02	0.364	
Familiarity	–0.04	0.308		–0.02	0.054	.
Semantic size	0.09	0.007	**	0.03	0.003	**
Gender association	0.00	0.949		–0.01	0.114	
Auditory strength	0.01	0.762		0.03	0.012	*
Gustatory strength	0.00	0.913		–0.04	< .001	***

[5] The largest adjusted GVIF in the imputed models was observed for *concreteness*. Sensitivity models excluding *concreteness* yielded broadly similar results; full outputs are provided in the OSF supplementary materials.

Haptic strength	0.06	0.148		0.06	< .001	***
Interoceptive strength	0.14	0.001	**	0.09	< .001	***
Olfactory strength	–0.08	0.025	*	–0.04	0.001	**
Visual strength	–0.02	0.668		–0.04	0.005	**
Max perceptual strength	0.04	0.615		0.10	< .001	***
Perceptual exclusivity	0.01	0.888		0.04	0.007	**
Cognition rating	–0.06	0.178		–0.13	< .001	***
Phonemic surprisal	–0.01	0.569		0.03	< .001	***
Word length	–0.06	0.200		–0.06	< .001	***
Etymology: prosaic	–0.81	< .001	***	–0.65	< .001	***
POS: Adjective	0.39	< .001	***	0.22	< .001	***
POS: Verb	0.24	0.003	**	0.20	< .001	***
POS: Adverb	0.30	0.576		–0.23	< .001	***
POS: Function word	0.39	0.611		–0.22	< .001	***
POS: Interjection	—	—		0.86	< .001	***

Iconicity variability (SD)

Raw data analysis In the raw-data analysis, the regression model accounted for a substantial proportion of variance in iconicity variability (adjusted R^2 = .19). Several predictors showed reliable associations. *Age of acquisition* was negatively associated with variability (β = −0.04, p = .028), whereas *familiarity* showed a positive association (β = 0.04, p = .038). Among perceptual predictors, *auditory strength* was positively associated with variability (β = 0.04, p = .021), while *maximum perceptual strength* showed a negative effect (β = −0.08, p = .017). *Semantic neighborhood density* showed a marginal negative association with variability (β = −0.04, p = .10). No reliable effects were observed for affective dimensions (*valence*, *arousal*, *dominance*), *word frequency*, *word length*, *phonological surprisal*, *etymology*, or *part-of-speech* categories.

Imputed data analysis In the full-coverage imputed analysis, the model accounted for 22% of the variance in iconicity variability. Several additional associations emerged in this analysis, which we interpret as extensions beyond the complete-case results rather than as direct consequences of imputation. Lexical frequency showed a positive association with iconicity variability (*word frequency*: β = 0.02, p < .001), as did *funniness* (β = 0.01, p = .012), *familiarity* (β = 0.02, p < .001), and *imageability* (β = 0.03, p < .001). *Word length* was also positively associated with variability (β = 0.02, p < .001), whereas *age of acquisition* showed a negative effect (β = −0.01, p = .002). Additional effects emerged for conceptual and distributional predictors. *Cogn*itive *rating* (β = −0.01, p = .012) and *semantic neighborhood density* (ARC; β = −0.02, p < .001) were negatively associated with variability, whereas *phonological surprisal* showed a positive effect (β = 0.01, p = .001). Among affective dimensions, *dominance* was negatively associated with variability (β = −0.02, p < .001). With nouns as the reference category, *part-of-speech* effects newly emerged for adjectives (β = −0.03, p < .001), adverbs (β = −0.04, p = .015), function words (β = 0.11, p < .001), and verbs (β = −0.06, p < .001).

Discussion

The present analyses largely replicate earlier findings from large-scale studies of English iconicity ratings (Winter et al., 2017, 2024), while extending them with additional predictors not previously examined. Consistent with prior work, words rated as more iconic were judged to be *funnier*, supporting the link between iconicity and playful aspects of language (Dingemanse & Thompson, 2020). Iconicity ratings were negatively associated with *semantic neighborhood density* (ARC), indicating that words with more distinctive semantic neighborhoods tend to be perceived as more iconic (Sidhu & Pexman, 2018). We also replicated robust negative relationships between iconicity and *age of acquisition* (Perry et al., 2015; Winter et al., 2017, 2024; Vinson et al., 2008), *word frequency* (Perry et al., 2015, 2018; Winter et al., 2017, 2024), and *log letter frequency* (Winter et al., 2024), suggesting that iconic words are learned earlier, occur less frequently, and exhibit more marked phonological structure. *Part-of-speech* effects were likewise consistent with previous reports, with nouns showing lower iconicity ratings than adjectives (Perry et al., 2015, 2018; Winter et al., 2017, 2024). In addition, the negative association between iconicity and *olfactory strength* replicated the pattern reported by Winter et al. (2017). Together, these replications provide a strong baseline validation of the present analyses and confirm that the raw-data results align closely with established findings in the literature. *Etymological imitativeness* emerged as a robust predictor of iconicity, even when controlling for a wide range of lexical, semantic, and affective variables. This finding highlights the lasting contribution of historical form–meaning motivations to present-day perceptions of iconicity, beyond what can be captured by synchronic psycholinguistic measures alone.

In the full-coverage imputed analysis, several additional associations were observed. These included *word length* (negative) and multiple sensorimotor dimensions (*auditory* and *haptic*: positive; *visual*: negative), aligning with Winter et al. (2017). The negative association between *cognition ratings* and iconicity supports the findings of Lupyan and Winter (2018). The positive association between *phonemic*

surprisal and iconicity, as well as the *part-of-speech* effects for interjections, adverbs, and function words, are consistent with Winter et al. (2024). However, some predictors, including dominance and POS: adverb, differed in direction across the complete-case and imputed analyses. We therefore interpret these directionally unstable effects cautiously. Overall, the imputed results largely converged with earlier findings (Winter et al., 2017, 2024; Kilpatrick & Bundgaard-Nielsen, 2025), providing converging evidence for the validity of our imputation approach.

Several of these, including imputation-only effects, offer additional theoretical insight. The positive effect of *perceptual exclusivity* suggests that meanings grounded strongly in a single perceptual modality may promote clearer form–meaning mappings. *Semantic size* was also positively related to iconicity, suggesting a link to size-related sound symbolism. While both largeness and smallness are suggested to be iconically mappable based on previous work (Sapir, 1929), the possibility that size-related iconicity may be asymmetric has received little explicit attention. The results suggest that increases in magnitude may give rise to more robust or salient form–meaning correspondences than reductions, although this interpretation remains to be tested directly. Another finding that may at first glance appear counterintuitive is the positive association between interoceptive strength and iconicity. This may appear to conflict with observations that ideophones rarely target internal bodily states (Dingemanse, 2012). However, this apparent tension can be resolved by distinguishing between lexical classes and perceptual dimensions. While languages may lack large, dedicated ideophone inventories for interoceptive meanings, words that evoke internal sensations can still invite iconic interpretations through prosody, phonological structure, or analogical mapping. Interoceptive grounding may also contribute to perceived iconicity alongside other sensory domains (Connell et al., 2018).

Regarding inter-rater variability, several predictors showed systematic associations with variability. *Age of acquisition* was negatively associated with variability: early-acquired words showed greater divergence in iconicity ratings. In addition, *familiarity* was positively associated with variability: more familiar words elicited more divergent iconicity ratings. Taken together, these findings suggest that familiar or early-established words may invite more diverse individual associations, rather than necessarily producing greater agreement across speakers. Several predictors associated with experiential richness were linked to greater inter-rater variability. Higher *funniness*, *imageability*, and *phonemic surprisal* were associated with increased divergence in iconicity ratings, suggesting that humorous, image-rich, or phonologically unexpected words invite more subjective interpretation and multiple possible form–meaning mappings across individuals. In contrast, *maximum perceptual strength* was negatively associated with variability, indicating that words with strong and concentrated perceptual grounding tend to elicit more consistent iconicity ratings. Together, these patterns suggest that perceptual grounding and experiential or evaluative dimensions interact in complex ways in shaping individual differences in iconicity ratings.

Value of Imputation

Imputation has begun to be adopted in large-scale iconicity research (e.g., Dingemanse & Thompson, 2020), yet detailed diagnostics of imputation accuracy remain underreported. In the present study, imputation errors were generally small to moderate relative to the observed data ranges, corresponding to approximately 0.1%–10% of the range. In relative terms, *semantic neighborhood density* (9.8%) and *arousal* (8.6%) showed comparatively larger errors and should therefore be interpreted with additional caution. Because imputation relies on relationships among predictors, imputed-data results are best treated as conservative extensions of the complete-case analyses. Nevertheless, when accompanied by transparent diagnostics, imputation can help expand lexical coverage and generate new hypotheses about iconicity across the lexicon.

Conclusion and Future Directions

This study aimed to provide a comprehensive account of word-level iconicity by leveraging large-scale iconicity datasets, including analyses based on imputed data to maximize lexical coverage. Our results identified a broad range of psycholinguistic and lexical factors associated with iconicity, while also demonstrating that mean iconicity and inter-rater variability capture partially distinct aspects of iconic perception. These distinctions have important implications for language acquisition and processing. Iconic mappings that are perceived consistently across individuals—reflected in low inter-rater variability—may serve as especially reliable cues for learners, facilitating early form–meaning mapping and supporting bootstrapping in word learning. In contrast, iconicity that is highly subjective, even when salient on average, may place greater demands on individual experience or contextual support during acquisition and online processing.

Future research can build on these findings in several directions. First, replication across languages and modalities—including both spoken and signed languages—will be crucial for assessing the generality of the observed patterns. Second, individual-difference approaches may clarify which learner or speaker characteristics (e.g., personality, imagery ability, linguistic background) shape sensitivity to iconicity and contribute to variability in iconicity ratings. Third, further work is needed to identify which types of sound–meaning mappings are most readily perceived and most consistently shared across individuals or languages, and how such mappings interact with learning environments and communicative contexts. Together, these directions point toward a more nuanced understanding of iconicity as a multidimensional phenomenon shaped jointly by linguistic structure, perceptual grounding, and individual experience.

Acknowledgements

We are grateful to the three anonymous reviewers for their detailed and constructive comments on an earlier version of this paper, which helped us improve the clarity and framing of the study. Any remaining errors are our own. This work was supported by JSPS KAKENHI Grant Number 24KJ1308.

References

Bosworth, R. G., & Emmorey, K. (2010). Effects of iconicity and semantic relatedness on lexical access in American Sign Language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1573–1581. https://doi.org/10.1037/a0020934

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990. https://doi.org/10.3758/BRM.41.4.977.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*(4), 991-997. https://doi.org/10.3758/s13428–012–0190-4.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904-911. https://doi.org/10.3758/s13428–013–0403-5.

Caselli, N. K., & Pyers, J. E. (2017). The road to language learning is not entirely iconic: Iconicity, neighborhood density, and frequency facilitate acquisition of sign language. *Psychological Science, 28*(7), 979-987. https://doi.org/10.1177/0956797617700498.

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., & Emmorey, K. (2017). ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods, 49*(2), 784–801. https://doi.org/10.3758/s13428–016–0742–0.

Connell, L., Lynott, D., & Banks, B. (2018). Interoception: The forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1752), 20170143. https://doi.org/10.1098/rstb.2017.0143.

Corenblum, H. T., & Pexman, P. M. (2025). Cognition ratings for 8826 English words. *Behavior Research Methods*, 57(7), 181. https://doi.org/10.3758/s13428–025–02663-w

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., Zeller, J., Winter, B., & Perlman, M. (2021). Novel vocalizations are understood across cultures. *Scientific Reports*, 11(1), 10108. https://doi.org/10.1038/s41598–021-89445-4.

De Varda, A. G., Lamarra, T., Ravelli, A. A., Saponaro, C., Giustolisi, B., & Bolognesi, M. (2025). IconicITA: Iconicity ratings of the Italian affective lexicon. *PLOS One, 20*(12), e0337947. https://doi.org/10.1371/journal.pone.0337947.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass, 6*(10), 654-672. https://doi.org/10.1002/lnc3.361

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences, 19*(10), 603-615. https://doi.org/10.1016/j.tics.2015.07.013.

Dingemanse, M., & Thompson, B. (2020). Playful iconicity: Structural markedness underlies the relation between funniness and iconicity. *Language and Cognition, 12*(1), 203-224. https://doi.org/10.1017/langcog.2019.49.

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods, 50*(3), 1116-1124. https://doi.org/10.3758/s13428–017–0930-6.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. arXiv Preprint arXiv:2003.06505. https://doi.org/10.48550/arXiv.2003.06505.

Fay, N., Ellison, M., & Garrod, S. (2014). Iconicity: From sign to system in human communication and language. *Pragmatics & Cognition, 22*(2), 244-263. https://doi.org/10.1075/pc.22.2.05fay

Ferrara, L., & Hodge, G. (2018). Language as description, indication, and depiction. *Frontiers in Psychology, 9*, 716. https://doi.org/10.3389/fpsyg.2018.00716.

Hinojosa, J. A., Haro, J., Magallares, S., Duñabeitia, J. A., & Ferré, P. (2021). Iconicity ratings for 10,995 Spanish words and their relationship with psycholinguistic variables. *Behavior Research Methods, 53*(3), 1262-1275. https://doi.org/10.3758/s13428–020–01496-z.

Iida, H., & Akita, K. (2023). Perceptual strength norms for 510 Japanese words, including ideophones: A comparative study with English. *Proceedings of the Annual Meeting of the Cognitive Science Society, 45*. https://escholarship.org/uc/item/1cg0835m.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1651), 20130298.

Kilpatrick, A., & Bundgaard-Nielsen, R. L. (2025). Exploring the dynamics of Shannon's information and iconicity in language processing and lexeme evolution. *PLOS One, 20*(4), e0321294. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0321294.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978-990. https://doi.org/10.3758/s13428–012–0210-4.

Lupyan, G., & Winter, B. (2018). Language is more abstract than you think, or, why aren't languages more iconic? Philosophical Transactions of the Royal Society B:

Biological Sciences, 373(1752), 20170137. https://doi.org/10.1098/rstb.2017.0137

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods, 52*(3), 1271-1291. https://doi.org/10.3758/s13428–019–01316-z.

Macuch Silva, V., Holler, J., Ozyurek, A., & Roberts, S. G. (2020). Multimodality and the origin of a novel communication system in face-to-face interaction. *Royal Society Open Science, 7*(1), 182056. https://doi.org/10.1098/rsos.182056

Massaro, D. W., & Perlman, M. (2017). Quantifying iconicity's contribution during language acquisition: Implications for vocabulary learning. *Frontiers in Communication, 2*, 4. https://doi.org/10.3389/fcomm.2017.00004.

McLean, B., Dunn, M., & Dingemanse, M. (2023). Two measures are better than one: Combining iconicity ratings and guessing experiments for a more nuanced picture of iconicity in the lexicon. *Language and Cognition, 15*(4), 716-739. https://doi.org/10.1017/langcog.2023.9.

Motamedi, Y., Little, H., Nielsen, A., & Sulik, J. (2019). The iconicity toolbox: empirical approaches to measuring iconicity. Language and Cognition, 11(2), 188-207. https://doi.org/10.1017/langcog.2019.14.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review 13*(1), 45-52. https://doi.org/10.3758/BF03193811

Occhino, C., Anible, B., Wilkinson, E., & Morford, J. P. (2017). Iconicity is in the eye of the beholder: How language experience affects perceived iconicity. *Gesture, 16*(1), 100-126. https://doi.org/10.1075/gest.16.1.04occ.f201

Ortega, G., Schiefner, A., Lazarus, N., & Perniss, P. (2025). A lexical database of British Sign Language (BSL) and German Sign Language (DGS): Iconicity ratings, iconic strategies, and concreteness norms. *Behavior Research Methods, 57*, 139. https://doi.org/10.3758/s13428–025–02660-z.

Ortega, G. (2017). Iconicity and sign lexical acquisition: A review. *Frontiers in Psychology, 8*, 1280. https://doi.org/10.3389/fpsyg.2017.01280.

Ortega, G., Sümer, B., & Özyürek, A. (2017). Type of iconicity matters in the vocabulary development of signing children. Developmental Psychology, 53(2), 345–357. https://doi.org/10.1037/dev0000161.

Perlman, M., & Lupyan, G. (2018). People Can Create Iconic Vocalizations to Communicate Various Meanings to Naïve Listeners. *Scientific Reports, 8*(1), 2634. https://doi.org/10.1038/s41598–018-20961-6.

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society Open Science, 2*(8), 150152. https://doi.org/10.1098/rsos.150152.

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Psychology, 1*, 227. https://doi.org/10.3389/fpsyg.2010.00227.

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: from a world of experience to the experience of language. Philosophical *Transactions of the Royal Society B: Biological Sciences, 369*(1651), 20130300. https://doi.org/10.1098/rstb.2013.0300.

Perry, L. K., Custode, S. A., Fasano, R. M., Gonzalez, B. M., & Savy, J. D. (2021). What is the buzz about iconicity? How iconicity in caregiver speech supports children's word learning. *Cognitive Science, 45*(4), e12976. https://doi.org/10.1111/cogs.12976.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLOS ONE, 10*(9), e0137147. https://doi.org/10.1371/journal.pone.0137147.

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2018). Iconicity in the speech of children and adults. *Developmental Science, 21*(3), e12572. https://doi.org/10.1111/desc.12572.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*(3), 225–239. https://doi.org/10.1037/h0070931

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods, 51*(3), 1258-1270. https://doi.org/10.3758/s13428–018-1099-3.

Sehyr, Z. S., & Emmorey, K. (2019). The perceived mapping between form and meaning in American Sign Language depends on linguistic knowledge and task: Evidence from iconicity and transparency judgments. Language and *Cognition, 11*(2), 208-234. https://doi.org/10.1017/langcog.2019.18.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods, 42*(2), 393-413. https://doi.org/10.3758/BRM.42.2.393

Sidhu, D. M., & Pexman, P. M. (2018). Lonely sensational icons: semantic neighbourhood density, sensory experience and iconicity. *Language, Cognition and Neuroscience, 33*(1), 25-31. https://doi.org/10.1080/23273798.2017.1358379.

Sidhu, D. M., Vigliocco, G., & Pexman, P. M. (2020). Effects of iconicity in lexical decision. *Language and Cognition, 12*(1), 164-181. https://doi.org/10.1017/langcog.2019.36.

Thompson, A. L., Akita, K., & Do, Y. (2020). Iconicity ratings across the Japanese lexicon: A comparative study with English. *Linguistics Vanguard, 6*(1), 20190088. https://doi.org/10.1515/lingvan-2019–0088.

Thompson, R. L. (2011). Iconicity in language processing and acquisition: What signed languages reveal. *Language and Linguistics Compass, 5*(9), 603-616. https://doi.org/10.1111/j.1749-818X.2011.00301.x.

Thompson, R. L., Vinson, D. P., & Vigliocco, G. (2010). The link between form and meaning in British sign language: effects of iconicity for phonological decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(4), 1017. https://doi.org/10.1037/a0019339.

Thompson, R. L., Vinson, D. P., Woll, B., & Vigliocco, G. (2012). The road to language learning is iconic: Evidence from British Sign Language. *Psychological Science, 23*(12), 1443-1448. https://doi.org/10.1177/0956797612459763.

Van Hoey, T., Yu, X., Pan, T. L., & Do, Y. (2024). What ratings and corpus data reveal about the vividness of Mandarin ABB words. *Language and Cognition, 16*(4), 1674-1696. https://doi.org/10.1017/langcog.2024.22.

Vinson, D. P., Cormier, K., Denmark, T., Schembri, A., & Vigliocco, G. (2008). The British Sign Language (BSL) norms for age of acquisition, familiarity, and iconicity. *Behavior Research Methods, 40*(4), 1079-1087. https://doi.org/10.3758/BRM.40.4.1079.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191-1207. https://doi.org/10.3758/s13428–012–0314-x

Winter, B., & Perlman, M. (2021). Iconicity ratings really do measure iconicity, and they open a new window onto the nature of language. *Linguistics Vanguard, 7*(1), 20200135. https://doi.org/10.1515/lingvan-2020–0135.

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic? *Interaction Studies, 18*(3), 443–464. https://doi.org/10.1075/is.18.3.07win.

Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2024). Iconicity ratings for 14,000+ English words. *Behavior Research Methods, 56*(3), 1640-1655. https://doi.org/10.3758/s13428–023–02112-6.

Winter, B., Woodin, G., & Perlman, M. (2026). Defining iconicity for the cognitive sciences. In O. Fischer, K. Akita, & P. Perniss (Eds.), *The Oxford handbook of iconicity in language* (pp. 138–150). Oxford University Press.