Iconicity ratings for 14,000+ English words

Winter, Bodo; Perlman, Marcus; Lupyan, Gary; Perry, Lynn; Dingemanse, Mark

DOI:
10.3758/s13428-023-02112-6

Iconicity ratings for 14,000+ English words

Bodo Winter[1], Marcus Perlman[1], Lynn K. Perry[2],
Gary Lupyan[3], Mark Dingemanse[4]

[1] Dept. of English Language & Linguistics, University of Birmingham, UK
[2] Department of Psychology, University of Miami, Florida
[3] Department of Psychology, University of Wisconsin, Madison
[4] Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

Abstract
Iconic words and signs are characterized by a perceived resemblance between aspects of their form and aspects of their meaning. For example, in English, iconic words include *peep* and *crash*, which mimic the sounds they denote, and *wiggle* and *zigzag*, which mimic motion. As a semiotic property of words and signs, iconicity has been demonstrated to play a role in word learning, language processing, and language evolution. This paper presents the results of a large-scale norming study for more than 14,000 English words conducted with over 1,400 American English speakers. We demonstrate the utility of these ratings by replicating a number of existing findings showing that iconicity ratings are related to age-of-acquisition, sensory modality, semantic neighborhood density, structural markedness, and playfulness. We discuss possible use cases and limitations of the rating dataset, which is made publicly available.

Key words: sound symbolism; arbitrariness of the sign; ideophones; onomatopoeia; crossmodal correspondence; lexicon

1. Introduction

Spoken words can sound like what they mean. This is the case for many onomatopoeic words that mimic qualities of sound such as the English words *bang* and *hiss*. Beyond sound, iconic words also mimic other sensory qualities, such as manner of movement as in the English *twirl* and *wiggle*, textures as in *mush* and *crispy*, visual events as in *flash* and *twinkle*, or size as in *teeny* and *humongous*. These words are iconic: aspects of their form are perceived to resemble aspects of their meaning.

Growing evidence now makes clear that iconicity is a foundational property of all human languages, spoken and signed (Dingemanse et al., 2015; Ferrara & Hodge, 2018; Perniss et al., 2010). Studies show that iconicity plays an active role in word learning (Imai & Kita, 2014; Ortega, 2017; Perniss & Vigliocco, 2014; Perry et al., 2015, 2018, 2021; R. L. Thompson et al., 2012), language processing (Bosworth & Emmorey, 2010; Sidhu et al., 2020; R. L. Thompson et al., 2010; Vinson et al., 2015), and language evolution (Ćwiek et al., 2021; Fay et al., 2014; Macuch Silva et al., 2020; Perlman et al., 2015; Verhoef et al., 2016).

In signed languages, iconicity is clearly evident (Cuxac, 1993). Although its importance has been historically downplayed (see discussion in Wilcox, 2004), it has been suggested for various signed languages that as many as half or three-quarters of signs appear to have iconic origins (Bellugi & Klima, 1975, 1975, 1976; Emmorey, 2014; Pizzuto & Volterra, 2000). For spoken languages, it has often been assumed that iconicity is confined to a small set of onomatopoeias. Contrary to this view, large-scale cross-linguistic analyses of lexicons find iconic form-meaning correspondences in basic vocabulary items (Blasi et al., 2016; Johansson et al., 2019; Joo, 2020; Wichmann et al., 2010), deictic terms (Johansson & Zlatev, 2013), color terms (Johansson et al., 2020), and texture words (Winter, Sóskuthy, et al., 2021). Within English, iconic sound-meaning correspondences have been established for size adjectives (Winter & Perlman, 2021b), touch adjectives (Winter, Sóskuthy, et al., 2021), and the visual shape of object nouns (Sidhu et al., 2021). Thus, iconicity can be found in many parts of the lexicons of natural languages, both signed and spoken.

Research on iconicity is characterized by methodological diversity (Motamedi et al., 2019), with different methods tapping into distinct, complementary aspects of the phenomenon (Dingemanse et al., 2020). The current study investigates iconicity through the lens of native speaker intuitions by asking raters to judge how much a word 'sounds like what it means' (Perry et al., 2015; Winter & Perlman, 2021a). This method was adopted from earlier studies that collected iconicity ratings for signed vocabularies (Grote, 2013; Lieberth & Gamble, 1991; R. L. Thompson et al., 2012; Vinson et al., 2008). Here, we present a new dataset of iconicity ratings for 14,776 English words. The ratings are freely available and can be downloaded in the following Open Science Framework (OSF) repository: https://osf.io/qvw6u/.

We start by characterizing in more detail what we mean by iconicity (Section 2.1), followed by a review of findings that have been obtained with earlier, smaller iconicity rating datasets (Section 2.2). Section 3 describes the methods used to collect

and analyze the ratings. Following this, the results report descriptive statistics of the ratings (Section 4.1), correlations with previous iconicity ratings (Section 4.2), and a set of replications of findings previously obtained with smaller iconicity rating datasets (Section 4.3). We conclude by discussing some limitations of the use of ratings as a method for the study of iconicity, as well as some key avenues for further research (Section 5).

2. Background

2.1. Defining and measuring iconicity

Researchers generally agree that iconicity refers to a quality of "resemblance" between the form and meaning of a signal. Yet, beyond this, iconicity is a field characterized by considerable diversity in basic concepts and terminology (Ahlner & Zlatev, 2010). Elsen (2017, p. 491) notes the lack of "a generally accepted definition," and Flaksman (2017, p. 18) remarks that the field is "still in need of a clear, established terminology."

One prominent confusion arises from the conflation of iconicity with systematicity. Dingemanse et al. (2015, p. 604) define systematicity as "a statistical relationship between the patterns of sound for a group of words and their usage", which may, or may not, be iconic. The sequence *gl-*, for example, occurs in many different words denoting shiny visual things like *glimmer, glitter, glitz* and *glisten* (Bergen, 2004; Bolinger, 1940, 1950; Firth, 1935; Marchand, 1959). This particular recurrence of form and meaning, called a 'phonestheme', is an example of systematicity, but in this case, the specific way form is linked to meaning does not appear to be based on any recognizable resemblance. That is, it is not obvious how *gl-* could be said to be iconic in the sense that it 'resembles' shiny visual things. In contrast, there are also many phonesthemes that are clearly iconic (Käsmann, 1992; Kwon, 2017; Kwon & Round, 2015), such as *cl-*, found in the onset of some onomatopoeias (e.g., *click, clonk, clack*), in which it mimics the abrupt onset of the sound to which it refers (Rhodes, 1994). Thus, phonesthemes nicely exemplify how systematicity is orthogonal to iconicity (Nielsen & Dingemanse, 2020; Nölle et al., 2018): systematic form-meaning correspondences in the lexicon can be iconic or non-iconic.

Systematicity is a quantitatively verifiable property of the lexicon of a language that is typically studied by performing statistical investigations of form-meaning mappings across a large number of vocabulary items (Monaghan et al., 2014; Sidhu et al., 2021; Winter, Sóskuthy, et al., 2021; Winter & Perlman, 2021b). In contrast to this, iconicity is often characterized as fundamentally subjective and dependent on interpretative processes, as is the case in signed language research (Cuxac, 1993; Wilcox, 2004), where it has been stated that iconicity is in "in the eye of the beholder" (Occhino et al., 2017). Several spoken language linguists (Jakobson & Waugh, 1979; Waugh, 1993, p. 73; Diffloth, 1994; Nuckolls, 2000) and literature scholars (Bredin, 1996; Hrushovski, 1980) too have emphasized the fluid and subjective nature of iconicity.

The idea that iconicity is subjective is supported by experimental studies demonstrating that the perception of iconicity varies as a function of individuals, tasks, and context. In signed languages, experiments have shown that the same sign has different iconic associations for different individuals (Occhino et al., 2017; Sehyr & Emmorey, 2019). For spoken languages, context dependence is demonstrated by experiments which show that the same speech sound can be perceived to resemble many different meanings depending on the task in which they are interpreted (French, 1977; Lockwood & Dingemanse, 2015; Winter et al., 2019). For example, the high-front vowel /i/ can be mentally associated not only with small size (Haynie et al., 2014; Newman, 1933; Sapir, 1929; Winter & Perlman, 2021b), but also with bitterness (Bankieris & Simner, 2014), angular shapes (O'Boyle & Tarte, 1980; Tarte, 1974), and brightness (Marks, 1974, 1982, 1989; Newman, 1933).

Heise (1966) described this aspect of iconicity as the "polysemy" of iconic meanings; Werner and Kaplan (1963) call it "plurisignificance" (see also Sidhu & Pexman, 2018a), and we have called it "pluripotentiality" (Winter et al., 2019; Winter, Oh, et al., 2021). What /i/ and other sounds 'mean' then is not invariable and static. The same way that different tasks can tap into different iconic associations of the same sound, words provide a semantic context that restricts pluripotentiality, narrowing down the range of latent iconic associations. For example, the phoneme /i/ is associated with angularity when embedded in pseudowords like *kiki* in the context of a psycholinguistics experiment (Bremner et al., 2013; Köhler, 1929; Ramachandran & Hubbard, 2001), but with small size when occurring in English words such as *teeny* and *meagre* (Winter & Perlman, 2021b). Orr (1944, p. 2) already noted that "it is the words and their setting which awaken the expressive possibilities latent in the sounds, and not the sounds which confer expressiveness to the words."

Taken together, our review of the literature leads us to define iconicity as follows:

> A signal in any medium or modality, such as a word, sign or gesture, is iconic to the extent that language users produce or perceive it through a sense of resemblance between some aspect of its form and some aspect of its meaning.

The distinct components of this definition are worth unpacking. First, our definition recognizes that iconicity is a modality-independent notion (Perniss et al., 2010). Second, the definition refers to a *perceived* resemblance in line with the subjective and interpretative nature of iconicity (Occhino et al., 2017; Wilcox, 2004). From this also follows a graded notion of iconicity, as we can perceive a form to be *more or less* similar to its meaning (Waugh, 1993, 1994). Third, the definition speaks of *some aspect of form* and *some aspect of meaning*, thereby recognizing that iconicity is always selective (Clark & Gerrig, 1990; Hassemer & Winter, 2018), i.e., iconic expressions always partial out specific sub-aspects of a phenomenon; words or signs don't mimic meanings in their totality.

How does the idea that iconicity is a subjective process of construal relate to research that investigates observable patterns of iconicity in the lexicon, such as the association between particular phonemes and particular meanings in the world's languages (Blasi et al., 2016; Johansson et al., 2019; Joo, 2020)? We think of these externally visible manifestations of iconicity as the imprint that people's iconic intuitions leave on the lexicon (cf. Taylor & Taylor, 1965). If enough people, for example, share the intuition that high-front vowels sound 'small' (Huang et al., 1969; Knoeferle et al., 2017; Newman, 1933; Sapir, 1929), these shared intuitions can become manifested as statistical regularities in the lexicon, such as English size adjectives referring to smallness being more likely to feature high-front vowels (e.g., *tiny, meagre, little, itsy-bitsy, mini*, see Winter & Perlman, 2021b). Such clusters of systematicity within the lexicon emerge because iconic intuitions have the power to shape the cultural evolution of the lexicon (Johansson et al., 2021; Vinson et al., 2021). Likewise, when iconic intuitions are shared across speakers or signers from different cultural backgrounds, universal patterns of form-meaning association emerge that can be captured via typological studies (Blasi et al., 2016; Haynie et al., 2014; Johansson et al., 2019, 2020; Johansson & Zlatev, 2013; Joo, 2020; Winter, Sóskuthy, et al., 2021).

2.2. Iconicity ratings

Several recent studies have measured the iconicity of words and signs by asking language users to rate them for how iconic they perceive them to be (Hinojosa et al., 2020; Motamedi et al., 2019; Perry et al., 2015; A. L. Thompson et al., 2020; Vinson et al., 2008; Winter & Perlman, 2021a). As this measure relies on the intuitions of language users, iconicity ratings, more than other methods, tap into the subjective dimension of the phenomenon, the extent to which language users *think* linguistic forms resemble their meanings. Large iconicity rating studies have been conducted for signed languages, including British Sign Language (Vinson et al., 2008) and American Sign Language (ASL, Caselli et al., 2017), and also in spoken languages, including English (Perry et al., 2015, 2018; Winter et al., 2017), Spanish (Hinojosa et al., 2020), and Japanese (A. L. Thompson et al., 2020). These studies have produced a number of findings for both signed and spoken languages, reviewed in Winter and Perlman (2021a), that we aim to replicate here for English.

First, for English and British Sign Language it has been shown that iconicity is higher for early learned words and signs (Massaro & Perlman, 2017; Perry et al., 2015, 2018; R. L. Thompson et al., 2012; Vinson et al., 2008), in line with the idea that iconicity may be helpful in word learning (Yoshida, 2004; Imai & Kita, 2014; Ortega, 2017; Nielsen & Dingemanse, 2020). Second, iconicity ratings in English, Spanish, and American Sign Language correlate with sensory experience ratings (Hinojosa et al., 2020; Perlman et al., 2018; Sidhu & Pexman, 2018b; Winter et al., 2017), i.e., concrete, perceptual concepts are more prone to being expressed iconically than abstract ones devoid of perceptual content (see also Lupyan & Winter, 2018).

Third, iconicity ratings in English correlate with humor ratings (Dingemanse & Thompson, 2020); for example, English words such as *smooch, waddle, pop, oink,* and *zigzag* are rated to be both iconic and funny. Fourth, in English but not American Sign Language, iconicity ratings are anti-correlated with the density of semantic neighborhoods (Sidhu & Pexman, 2018b; B. Thompson et al., 2020), which is generally explained as a result of ambiguity avoidance: if iconicity is associated with a cluster of systematicity in the lexicon, words or signs with similar meanings will tend to have similar forms, potentially leading to confusion (Gasser, 2004; Monaghan et al., 2012; B. Thompson et al., 2020). ASL may be an exception to this pattern because there are more degrees of freedom for iconic expression, which may help to reduce overlap in forms for words with similar meanings (B. Thompson et al., 2020).

Fifth, English iconicity ratings correlate with measures of structural markedness (Dingemanse & Thompson, 2020), which means that iconic words stand out from other words. This has been shown in terms of a) the presence or absence of complex onsets, which are more likely to occur in iconic words (e.g., *bleep, crunch, flap, flick, prick, sniff*), and b) log letter frequency, a coarse indicator of orthographic improbability. Sixth, it has been found that English iconicity ratings are negatively correlated with word frequency when these word frequencies are taken from adult corpora (Perry et al., 2015), but not when they are taken from child-directed speech (Perry et al., 2018). That is, there is an empirically demonstrated tendency for adults to use iconic words *less* often, but to use them *more* often when talking to young children acquiring language. Seventh and finally, iconicity ratings in English, Spanish, and Japanese differ across different parts of speech (Hinojosa et al., 2020; Perry et al., 2015, 2018; Winter et al., 2017). While there is some cross-linguistic variation about the overall ranking of different parts of speech, onomatopoeias and interjections generally receive higher ratings than verbs and adjectives, with nouns generally being rated least iconic.

In this paper, we use a much more extensive set of English iconicity ratings to replicate these findings in one simultaneous analysis. Besides expanding the scope of the words covered, thereby leading to more general results, this is the first time all of these measures are combined in a single analysis, which means that we are now in a position to demonstrate that these results hold when controlling for each other. These iconicity ratings also have uses beyond correlation studies with other rating scales. For example, iconicity ratings can aid in the selection of stimuli for psycholinguistic experiments on iconicity (e.g., Sidhu et al., 2020), or for the analysis of texts and discourse (Green & Perlman, 2022; Sidhu et al., 2022).

3. Methodology

3.1. Word list construction

Our list of English words was compiled with several criteria in mind. Specifically, we wanted to include 1) as many useful words as possible, i.e., words that would likely feature in experiments, and that are used fairly commonly in conversation and

writing, 2) words that overlapped with many different existing databases and show sufficient spread in terms of lexical variables of common interest (e.g., concreteness, frequency etc.), and 3) words that are known by a sufficiently high proportion of English speakers. We started by taking all the monomorphemic and bimorphemic words of the English Lexicon Project (Balota et al., 2007). We then added words trying to maximize overlap with humor ratings (Engelthaler & Hills, 2018), perceptual attribute ratings (Amsel et al., 2012; Medler et al., 2005), and touch ratings (Stadtlander & Murdoch, 2000) with consideration to planned future projects for which these additional rating scales are important. The word list also included the English glosses of the American Sign Language Lexicon (ASL-Lex) (Caselli et al., 2017), a long list of words containing phonesthemes (taken from Hutchins, 1998), a selection of words categorized as mass and count nouns (Kiss et al., 2016), and a list of verbs categorized for different lexical classes from Levin (1993). In addition, we included all the words from the earlier iconicity rating studies by Perry et al. (2018) and Winter et al. (2017) to facilitate comparison. The list was topped off with the 5% most and least concrete words based on norms of Brysbaert et al. (2014), as well as the 15% most and least positive words from the emotional valence norms of Warriner et al. (2013). The final word list presented to participants included 15,394 words.

3.3. Participants

The final dataset (see below for exclusion criteria) included 1,419 American English speakers (mean age = 30, SD=14, range = 18-88; 95.7% native English speakers, 51.8% female, 41.0% male, 0.5% other, 6.8% unreported) recruited via Amazon Mechanical Turk (55%) and the UW-Madison Psychology participant pool (43%). Participants recruited online were reimbursed $0.60 USD for rating 50 words and had the option to complete 1-2 additional 50-word lists for additional payment, a maximum of $1.90 for rating 150 words. Participants recruited from the UW-Madison participant pool were asked to rate 150 words in exchange for commensurate course credit.

3.2. Instructions

Participants were presented with words one at a time and asked to indicate how much they thought each word "sounds like" its meaning. They were asked to say the word out loud to themself, and to think about its meaning. The instructions for our previous iconicity ratings (Perry et al., 2018; Winter et al., 2017) were modelled after Vinson et al.'s (2008) iconicity ratings for British Sign Language (BSL): We gave examples of high or low iconicity words because the concept of iconicity is not necessarily known by laypeople (although people are generally familiar iconic phenomena like onomatopoeia and pantomime). We followed this approach by giving three examples each of words with low, medium, and high iconicity taken from our previous iconicity rating dataset (Perry et al., 2018; Winter et al., 2017). As critics of iconicity ratings have claimed that participants' ratings may be unduly influenced by onomatopoeias (A. L. Thompson et al., 2020), we selected examples of

highly iconic words that include not only onomatopoeias (*screech*), but also highly iconic words that are not as strongly tied to sound alone (*twirl* and *ooze*). The exact instructions and a sample trial are included in Appendix A.

3.3. Rating scale

We used a 7-point rating scale anchored at (1) "Not iconic at all" and (7) "Very iconic." The iconicity ratings we used in previous studies (Perry et al., 2015, 2018; Winter et al., 2017) used a scale ranging from -5 ("sounds like the opposite of what it means") to +5 ("sounds like what it means"), placing arbitrariness at the center of the scale at 0 ("does not sound like what it means or the opposite"). Here, we dispensed of the opposite end of the scale for several reasons. First, Perry et al. (2015) and Winter et al. (2017) already showed that the iconic end of the scale is used relatively little by participants, and Motamedi et al. (2019) observed that the lower end of the scale is used less consistently by participants as well (p. 197); it appears that participants do not have a clearly defined concept of what it means for a word to sound like the opposite of its meaning. For this reason, some studies that have used the previous English iconicity ratings have chosen to exclude words with negative iconicity ratings (cf. Sidhu & Pexman, 2018b). Second, our move away from using the negative end of the scale is also consistent with newer iconicity rating studies, such as Hinojosa et al. (2020) for Spanish, which employed a 1-7 Likert scale. Finally, using a 1-7 rather than a -5 to +5 scale makes our iconicity ratings more comparable to most other common rating scales used in large-scale norming studies (Brysbaert et al., 2014; Lynott et al., 2019; Warriner et al., 2013).

3.4. Procedure

Participants were given an option to skip a word if they did not know its meaning or pronunciation. Figure 1 shows a screenshot of a trial.

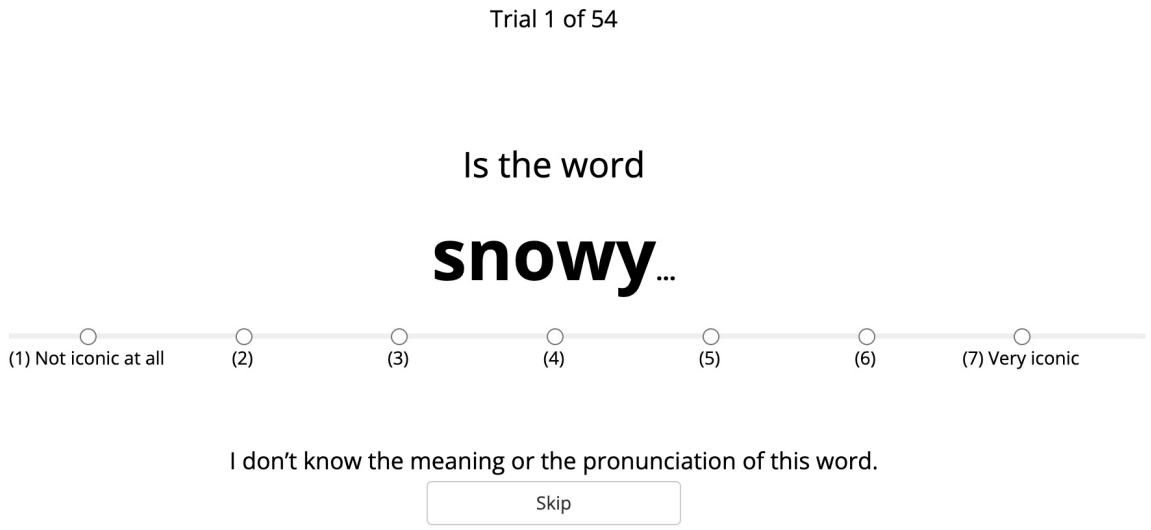

Figure 1. An example trial as presented to raters.

The rating task was implemented in jsPsych (De Leeuw, 2015) and deployed using Node.js in several rounds of data collection between 4/1/2020 and 9/1/2021, with additional data collection after review completed on 9/11/2022. We completed data collection until we reached at least 10 data points per word after exclusions (see below). The 10-ratings per word criterion was chosen following the recommendation of Motamedi et al. (2019), who show that the average iconicity rating of a word mostly stabilizes with about ~10 ratings, with increasing participant numbers yielding diminishing returns.

3.5. Data cleaning

Although data quality of Amazon Mechanical Turk has been independently validated many times for a wide range of behavioral findings (e.g., Rouse, 2015; Sprouse, 2011), there are well-known issues with crowdsourced data. We took several steps to maximize overall data quality. Table 1 details all the exclusions we undertook, and how many data points were excluded because of each criterion.

Criterion	**What's excluded**	**Excluded when**	**Number excluded**
attention checks	participant	failed ≥2 attention checks	112
response times	response	RT < 500ms	3,901
straightlining	participant	participants making more than 80% same responses	16
correlation with item averages	participant	participant's correlation with item averages was below Pearson's $r < 0.1$	70
word knowledge	word	known by less than 80% of all participants	618
number of ratings	word	fewer than 10 ratings per word	95

Table 1. Applied exclusion criteria[1]

Response times. The average response time was M = 3,847ms (SD = 61,054ms). We excluded response times faster than 500 ms. This threshold was chosen as it would be nearly impossible for participants to perform visual word recognition, make a comparative judgment involving deep processing of both phonology and semantics, and produce a keyboard/mouse response within this time span. Initial

[1] We additionally excluded data from people who began the task, but withdrew before completing at least 40 trials.

explorations also showed that response times below 500 ms were associated with people who disproportionately gave the same response ("straightliners"), suggesting that these data points are suspect. This lower threshold for response times led to the exclusion of 3,901 trials (2.2% of total trials).

Straightliners. Some survey respondents attempted to save time by giving identical or nearly identical responses to several survey items in a row (Y. Kim et al., 2019; Zhang & Conrad, 2014). We excluded 16 participants because more than 80% of their responses were the same value. Graphical exploration of straightlining as a function of response times also revealed that people who disproportionately gave the same response had unreasonably low average response times (see online repository).

Correlation with item averages. Following Warriner et al. (2013), we correlated each participant's individual ratings with the by-item averages from the remaining ratings. On average, participants were moderately well correlated with the averages for the items to which they responded (average Pearson's $r = 0.44$, $SD = 0.19$). 70 participants who had low or negative correlations with the item averages (Pearson's $r < 0.1$) were excluded.

Word knowledge. We excluded 618 words that were known by less than 80% of our participants, which included rare and obscure words such as *asbestos, bullion, vitrify, persiflage,* and *knave.*

Number of ratings. After several rounds of data collection, there were 95 words which failed to reach the 10-rating threshold due to a mixture of randomization and participants choosing the "I don't know the meaning or the pronunciation of this word" option. Average ratings for these words are not included in the final dataset, following the 10-rating threshold described by Motamedi et al. (2019).

Taken together, our exclusion criteria removed 20,871 individual ratings (11.5% of the original data). The remaining dataset is based on 161,057 individual ratings and includes 14,776 unique words. The file that contains the iconicity rating averages for each word ("iconicity_ratings_cleaned.csv") can be found in the OSF repository: https://osf.io/qvw6u/, together with the raw data before any exclusions ("iconicity_ratings_raw.csv"). We advise researchers to use the cleaned dataset.

3.6. Reliability Analyses

We assessed the reliability of the remaining ratings (after exclusion) using intraclass correlation coefficients (ICCs) implemented in the psych package version 2.1.9 (Revelle, 2021). Reliability of individual ratings was quite low (ICC2 = 0.13, 95% CI = 0.12-.13). Due to scale boundary effects, words with extreme ratings will tend to have less variation across raters, as has been discussed in the context of concreteness ratings (Pollock, 2018). Indeed, reliability for the iconicity ratings was higher for words outside the middle range by including only words with mean ratings of ≤3.5 or ≥4.5. Reliability was indeed higher for these more extreme iconicity ratings (ICC2 = 0.21, 95% CI = .21-.22), and increased further when examining only words outside the 3.25-4.75 range (ICC2 = .28, 95% CI = 0.28-0.29).

The above ICCs are measures of reliability of individual raters. These analyses show that individual raters vary considerably. However, because our dataset concerns *average* iconicity ratings, the more relevant reliability estimate is of a word's *average* rating. This is captured by the ICC2k measure which reflects the reliability of the *group* of raters (Shrout & Fleiss, 1979). ICC2k was 0.99, 95% CI=.99-.99 when the analysis included all words. It was within rounding error of 1.0 for the 9,282 words outside of the middle range.

To put these reliability estimates in perspective, we computed individual and average ratings for the widely used concreteness norms (Brysbaert et al., 2014). Reliability of individual estimates (ICC2) was 0.39, 95% CI = .38-.39 and increased to 0.51, 95% CI = .51-.52, when including the 43,935 with mean concreteness ratings ≤ 2.5 or ≥ 3.5 on a 5-point scale. The ICC of the averaged rating (ICC2k) was within rounding error of 1.[2]

3.7. Statistical analyses

All statistical analyses were conducted with R version 4.1.1 (R Core Team, 2019) and the "tidyverse" package 1.3.1 (Wickham et al., 2019) for data processing. Throughout the analysis, we use Bayesian regression models implemented with the "brms" package 2.16.2 (Bürkner, 2017).

Our main statistical model (Section 4.3) attempts to replicate results from four previous studies (Dingemanse & Thompson, 2020; Perry et al., 2018; Sidhu & Pexman, 2018b; Winter et al., 2017) in a single simultaneous regression analysis. For this, we regressed iconicity ratings on concreteness ratings (Brysbaert et al., 2014), sensory experience ratings (Juhasz & Yap, 2013), age-of-acquisition ratings (Kuperman et al., 2012), SUBTLEX corpus log frequencies (Brysbaert & New, 2009), humor ratings (Engelthaler & Hills, 2018), log-letter frequency as an indicator of structural markedness (following Dingemanse & Thompson, 2020), and average radius of co-occurrence (Shaoul & Westbury, 2010), a measure of semantic neighborhood density shown to correlate with rated iconicity (Sidhu & Pexman, 2018b). The only categorical predictor in this analysis was part-of-speech (Brysbaert et al., 2012).

For this model, we standardized all continuous predictors, which facilitated using the same weakly informative prior on all slope coefficients, for which we chose a normal distribution centered at zero with *SD* = 0.25. This prior assumes that 68% of all slope coefficients would fall in between -0.5 and +0.5 average ratings, and 95% of all slope coefficients fall in between -1 and +1. We chose this specific prior based on the largest effect observed in Winter et al. (2017), which, although using a different scale, is the most comparable analysis to what we report here. Due to it being centered at zero, the *Normal(0, 0.25)* prior on slope coefficients introduces "mild

[2] It is not inevitable that large enough groups of raters lead to ICC2k estimates of near 1. Random ratings predictably lead to averaged ICCs of 0. Adding increasing amounts of noise to the recorded ratings progressively lowers the ICC2k.

skepticism" into the model (McElreath, 2020), i.e., slightly biasing all coefficients towards zero. While it would be possible to come up with more specific priors for particular coefficients based on previous research, this would go against our goal of being able to compare the relative strength of each variable, as given by the new dataset. In addition, more specific priors for particular variables are hard to implement given that previous studies using the English ratings used a different scale (-5 to +5, rather than 1-7). We therefore decided to use the same prior for all slope coefficients. For the standard deviation, we chose a $Normal_{+}(0, 0.5)$ prior; for the intercept, we chose a prior focused at the midpoint of our scale: $Normal(4, 0.5)$.

Throughout the paper, we analyze item averages. Although this hides by-subject variation from the model, this is consistent with how rating studies are generally analyzed (e.g., Brysbaert et al., 2014; Kuperman, 2015; Warriner et al., 2013; Warriner & Kuperman, 2015). We do, however, incorporate variation across raters in a different way. Pollock (2018) emphasized that analyses of rating scales need to take the standard deviations across ratings into account. In all analyses below, words with low standard deviations (indicating more agreement between raters) contribute more to our overall results than words with high standard deviations. We achieve this by using the standard deviations (across raters for each word) as regression weights, which has also been shown to be effective for concreteness ratings, where models with regression weights penalizing high-SD words lead to higher model fit (Strik Lievers et al., 2021). Standard deviations were first rescaled so that 0 indicates the maximal standard deviation and 1 indicates the lowest standard deviation in the dataset. We subsequently renormalized these weights to have a mean of 1 (Gelman et al., 2020, p. 148).

As the item averages are reasonably well described by a normal distribution (see Figure 1a, below) we used a normal likelihood in all models below. Posterior predictive checks show that this is a reasonable assumption, although the model cannot simulate all patterns in the data. All models were estimated via MCMC with 4 chains (4,000 iterations, 2,000 warm-up). There were no divergent transitions and all chains mixed well (*Rhat* = 1.0 for all models). Analysis data and code can be found in the OSF repository: https://osf.io/qvw6u/

4. Results

4.1. Overview and descriptive statistics

Table 2 shows the ten most and least iconic words in the dataset, together with the corresponding item means and standard deviations. The most iconic words include onomatopoeia such as *oomph, clunk*, and *purr*. Non-onomatopoeic examples of words with high iconicity ratings include *wiggle, wobbly, puffy, crispy, zap, wring, crumbly, yucky, squash, cheesy, sniff, whiff, stink*, and *gloom*.

Most iconic	**Mean**	***SD***		**Least iconic**	**Mean**	**SD**
oomph	6.9	0.29		*how*	1.3	0.95

swish	6.9	0.30		*if*	1.3	0.48
wiggle	6.9	0.32		*partial*	1.3	0.68
clunk	6.8	0.42		*are*	1.4	0.67
creak	6.8	0.63		*gnome*	1.4	0.84
purr	6.8	0.42		*Rugby*	1.4	0.97
sigh	6.8	0.42		*shape*	1.4	0.90
squeak	6.8	0.42		*cerebellum*	1.5	0.97
woof	6.8	0.63		*incorruption*	1.5	0.67
bang	6.8	0.45		*ordain*	1.5	0.70

Table 2. The ten most and least iconic words for this dataset

The average iconicity rating was close to the middle of the 1-to-7 scale, with $M = 3.8$ ($SD = 0.9$). Figure 2a shows the distribution of iconicity ratings with a superimposed normal distribution with the same mean and SD. As can be seen in the figure, there were very few words with extremely low or extremely high iconicity ratings. Words with iconicity ratings in the middle of the scale had higher standard deviations (Figure 2b). Not surprisingly, raters agreed more with each other for words at the ends of the scale; see also Pollock (2018).

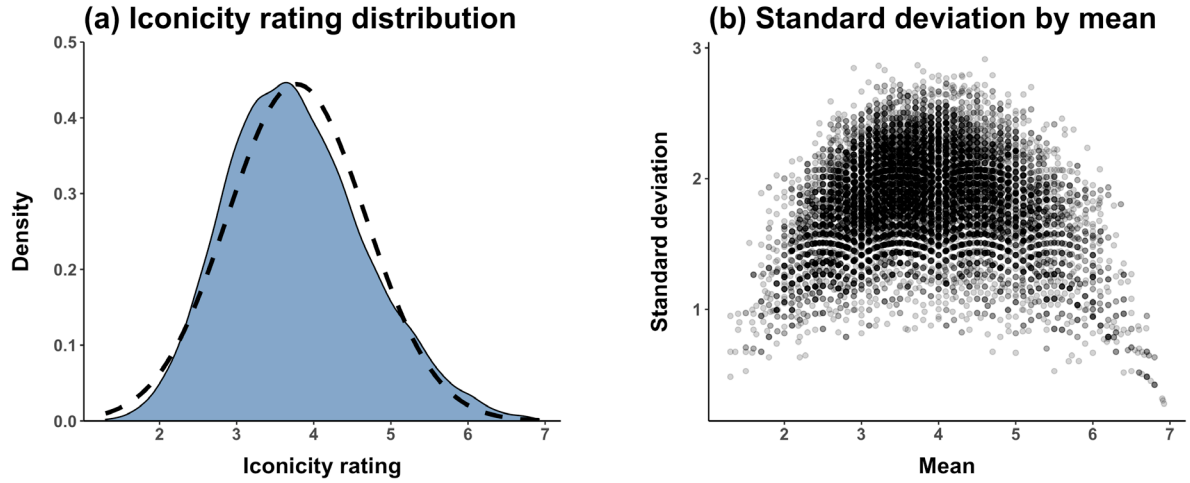

Figure 2. (a) A Kernel density plot of the distribution of average ratings; the dashed line indicates a normal distribution with the same mean and standard deviation; (b) standard deviations across raters (*y*-axis) as a function of average rating (x-axis), following Pollock (2018)

4.2. Correlation with existing iconicity rating datasets

To establish continuity with past research using iconicity ratings, we correlated our new ratings with ratings taken from prior studies. Perry et al. (2015) focused on 592 words that were rated for iconicity based on their written (experiment 1) or spoken form (experiment 2). Our ratings correlated with the written ratings from Perry et al. (2015), $r = 0.55$, 95% confidence interval: [0.49, 0.60]. They also correlated with the

spoken ratings from that study, but less so, r = 0.48, 95% CI: [0.42, 0.54]. The correlation was higher still (r = 0.63 [0.60, 0.65]) with the iconicity ratings for 3,000 English words from Perry et al. (2018) and Winter et al. (2017) (written presentation format only). One factor that may have lowered correlations in these comparisons is the fact that we used a different rating scale in our new study.

Dingemanse and Thompson (2020) used distributional semantics to impute iconicity values for words for which no ratings were available. We found a moderate correlation between our new ratings with the imputed ones, r = 0.50 [0.49, 0.52]. This correlation between contextually inferred iconicity and our new iconicity ratings is theoretically interesting in its own right as it shows that words of similar iconicity levels have similar distributional profiles in corpora and thereby shared semantic properties.

4.3. Replication of previous findings with iconicity rating datasets

In this section, we replicate the previous findings obtained with smaller iconicity rating datasets reviewed in Section 2.2 (see also Winter & Perlman, 2021a). The full model (with regression weights) described 29% of the variance, which is larger than the variance described by the model without regression weights (24%). On top of the conceptual considerations detailed in Pollock (2018) and above, this difference in model fit alone demonstrates the utility of incorporating standard deviations as regression weights (see also Strik Lievers et al., 2021).

Figure 3 shows all standardized coefficients with their 95% credible intervals at a glance. This figure excludes the categorical part-of-speech predictor, discussed below.

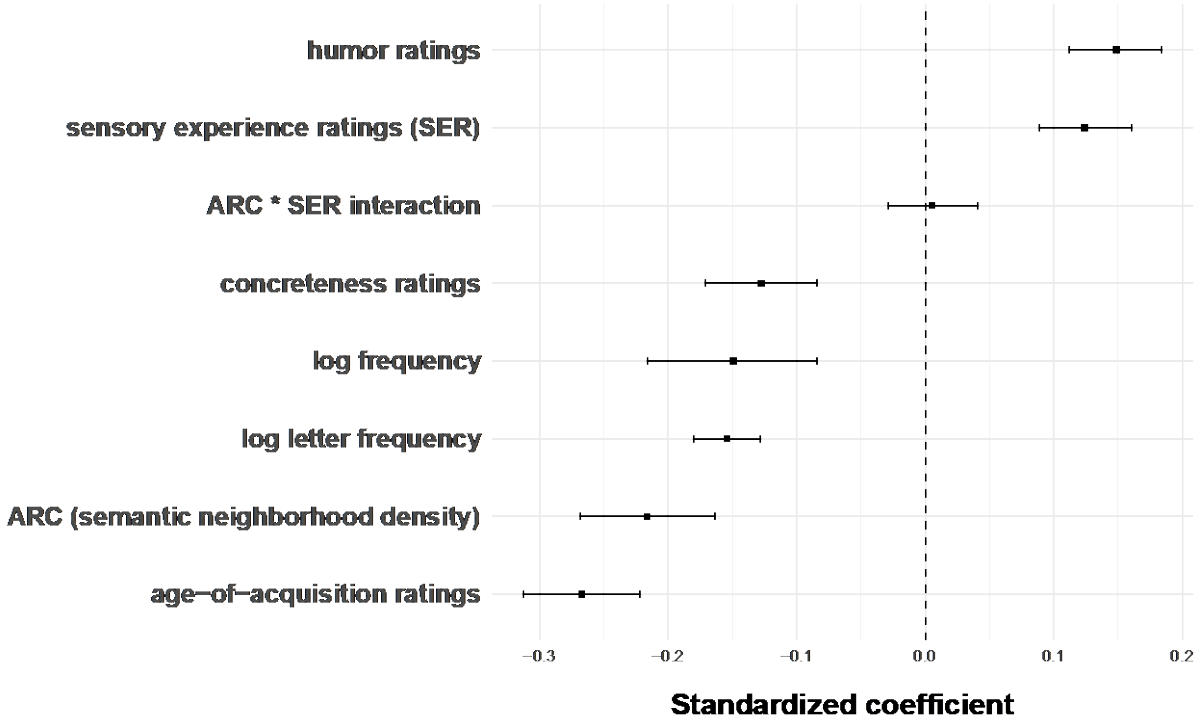

Figure 3. Coefficients and their associated 95% Bayesian credible intervals for all continuous predictors in the main multiple regression analysis (which also includes the categorical part-of-speech predictor, not shown here)

Humor ratings from Engelthaler and Hills (2018) were positively associated with iconicity ratings (posterior mean: +0.15, *SE* = 0.02), with a 95% credible interval that was far away from zero: [0.11, 0.18], thereby replicating Dingemanse and Thompson (2020).

Iconicity ratings were also correlated with sensory experience ratings from Juhasz and Yap (2013), with a coefficient (+0.12, *SE* = 0.02) that also does not overlap with zero: [0.09, 0.16], thereby replicating Winter et al. (2017) and Sidhu and Pexman (2018b). There was, however, a negative correlation with the concreteness ratings from Brysbaert et al. (2014): -0.13, *SE* = 0.02, [-0.17, -0.08], with more abstract words also being rated as *more* iconic. We discuss this somewhat counterintuitive finding in more detail below.[3]

As was also found by Sidhu and Pexman (2018b) for the previous ratings, the new iconicity ratings negatively correlate with semantic neighborhood density as measured by ARC (-0.22, *SE* = 0.03, 95% interval of coefficient: [-0.27, -0.16]). However, we failed to replicate the interaction they report between ARC and sensory experience ratings (~0.0, *SE* = 0.02, 95% interval: [-0.03, +0.04]).

Iconicity ratings correlated negatively with age-of-acquisition ratings from Kuperman et al. (2012) (-0.27, *SE* = 0.02, 95% interval: [-0.31, -0.22])—a replication of Perry et al.'s (2018) finding that earlier learned words tend to be more iconic.

Iconicity ratings were also negatively correlated with (log-transformed) letter frequency (-0.15, *SE* = 0.01, 95% interval: [-0.18, -0.13]). As discussed above, this is a coarse indicator of a word's orthographic probability that was previously used by Dingemanse and Thompson (2020) as a proxy for structural markedness.

Finally, words rated high in iconicity were less frequent on average (-0.15, *SE* = 0.03, 95% interval: [-0.22, -0.08]), as has been found in several previous studies for word frequency data from adult speakers (Perry et al., 2015, 2018; Winter et al., 2017).

Table 3 shows the descriptive averages for all parts-of-speech, which replicates the basic pattern established in previous iconicity rating studies on English (Perry et al., 2015, 2018; Winter et al., 2017). As can be seen in the table, interjections had the highest rated iconicity. It is worth noting that this relatively small group includes onomatopoeic interjections (e.g., *pop, quack, wham, blah*) but also interjections that depict emotional vocalizations (e.g., *yuck, ouch, ugh*). Verbs were higher in rated iconicity than nouns, with adjectives assuming an intermediate

[3] This result is unlikely driven by collinearity between sensory experience ratings and concreteness ratings. Variance inflation factors computed with the "car" package version 3.0.11 (Fox & Weisberg, 2018) suggest that there is little collinearity (all VIFs < 2). Moreover, dropping sensory experience ratings does not invert the sign, and neither does dropping any other predictor (e.g., frequency, AOA).

position. Adverbs (which largely include words with grammatical or discursive functions) and function words were lowest in rated iconicity, consistent with the observation that these word classes generally do not encode sensory perceptions and are also on average very abstract (Strik Lievers et al., 2021). To perform an omnibus test for the multi-level part-of-speech (POS) predictor, we performed leave-one-out cross-validation (LOO-CV), which indicated that the model without the POS predictor performed reliably worse in terms of predictive accuracy (elpd_diff = -73.4, *SE* = 14.5) than the model with this predictor. When the POS predictor was dropped from the model, the described variance of the overall model dropped from 29% to 26%.

Lexical category	N	Mean	*SD*
interjection	41	5.34	1.17
verb	2,770	3.97	0.95
adjective	3,054	3.79	0.87
noun	7,722	3.75	0.85
adverb	218	3.35	0.85
function words	184	3.20	0.79

Table 3. Iconicity rating means and standard deviations for each English part-of-speech (787 words not classified according to SUBTLEX POS tags)

6. Discussion

6.1. Summary of findings

We collected iconicity ratings from English speakers for more than 14,000 English words. With the exception of the interaction between semantic neighborhood density and sensory experience ratings reported in Sidhu and Pexman (2018b), all major results from previous rating studies replicated with the new norms (see Winter and Perlman 2021a for a review). In summary, we found that iconicity ratings were highest for sensory words (Sidhu & Pexman, 2018b; Winter et al., 2017), early acquired words (Perry et al., 2018), words that occupy sparse semantic neighborhoods (Sidhu & Pexman, 2018b), and words that are structurally marked and playful in nature (Dingemanse & Thompson, 2020). We also found that English interjections and verbs had higher iconicity ratings than adjectives, nouns, adverbs, and function words (Perry et al., 2015).

As was found previously, there also was a negative correlation between iconicity ratings and the concreteness ratings from Brysbaert et al. (2014). This finding is in need of some explanation, as it could be seen as superficially contradicting Lupyan and Winter's (2018) claim that abstract concepts are hard to express iconically. For the older iconicity rating dataset, Winter et al. (2017) already found that sensory experience ratings from Juhasz and Yap (2013) are more strongly associated with iconicity ratings than concreteness ratings. Several researchers have criticized the construct validity of concreteness ratings based on multiple arguments

(Connell & Lynott, 2012; Löhr, 2021; Strik Lievers et al., 2021; Winter, 2022). In particular, Connell and Lynott (2012) suggest that the concreteness ratings may be biased towards visual experience at the expense of other ways in which a concept can be accessible to the senses. Connected to this, Winter et al. (2017) found that highly visual concepts are, on average, not prone to iconic expression, perhaps due to the fact that many purely visual concepts such as color are hard to express iconically via speech alone. Another reason for iconicity ratings being negatively correlated with concreteness ratings may have to do with the fact that many abstract concepts receive high auditory ratings in sensory modality rating studies (Lynott et al., 2019; Lynott & Connell, 2013), with audition being one of the most iconic modalities for spoken languages (Perlman et al., 2018; Winter et al., 2017). For these combined reasons, when concreteness ratings are entered into a model together with sensory experience ratings, it actually comes as no surprise that concreteness is negatively associated with iconicity when sensory experience is held constant.

These replications expand on previous findings in two important ways. First, by covering far more words than previous studies, we are able to put the existing findings on a firmer quantitative footing and achieve more generalizable results. We are also able to show that previous results hold even when a different scale is used. Second, in the new analysis we were able to add all predictors simultaneously in the same regression model, something that is made possible by having more data for all predictors. Thus, our analysis demonstrates that results obtained by the individual studies (Dingemanse & Thompson, 2020; Perry et al., 2015, 2018; Sidhu & Pexman, 2018b; Winter et al., 2017) hold even when additional predictors are held constant. For example, words rated high in iconicity have low rated age-of-acquisition even when controlling for sensory experience, playfulness, structural markedness, etc. Taken together, these correlations make a strong case for the construct validity of the iconicity ratings (Winter & Perlman, 2021a), as this pattern of correlations is exactly what we would expect to see if the rating scale was actually measuring iconicity.

It is also worth highlighting that our replication study also expands on previous analyses by incorporating standard deviations. The fact that this is theoretically motivated (cf. Pollock, 2018) and also improved the fit of the model makes an important methodological point. Given that agreement between raters is not equal across the rating scale, analyses of iconicity ratings should take standard deviations into account. Notably, the use of standard deviations as regression weights has also been shown to increase model fit for other rating scales, such as concreteness ratings (Strik Lievers et al., 2021). Regression weights provide an easy way of incorporating disagreement between one's raters into one's analysis, and researchers using our iconicity ratings should consider this approach.

6.2. Correspondences with ideophone research

How do our findings for English words compare to other spoken languages? In this section, we draw an explicit connection between our findings in English and other lines of research focused on languages with large sets of explicitly imitative words,

variously called "ideophones", "mimetics" or "expressives" (Akita & Pardeshi, 2019; Dingemanse, 2019; F. E. Voeltz & Kilian-Hatz, 2001). For example, Japanese is reported to have thousands of ideophones such as *sarasara* (for smooth surfaces), *pikapika* (for bright and shiny sensations), or *zukizuki* (for throbbing pain). These depictive words often stand out from other words of a language by virtue of having unusual (i.e., 'marked') phonological, morphological, or syntactic patterns (Akita, 2009; Childs, 1994; F. E. Voeltz & Kilian-Hatz, 2001), and they often have a performative quality and tend to be associated with co-speech gestures (Nuckolls, 2020). Iconicity ratings of Japanese vocabulary confirm that native Japanese speakers judge ideophones to be more iconic than other words (A. L. Thompson et al., 2020).

Importantly, what constitutes an ideophone within a specific language can only be decided based on language-internal criteria, just as with other word classes (Dryer, 1997; Croft & van Lier, 2012; Dingemanse, 2019). The specific formal characteristics that make ideophones stand out from other words vary across languages (Childs, 1994). In contrast to languages such as Japanese for which there are relatively clear formal criteria for determining whether a word is an ideophone (Akita, 2009), this distinction appears to be less clear in standard European languages such as English. Some analyses have classified English onomatopoeic words such as *boom, tweet, zap,* and *poof* as ideophones (e.g., A. L. Thompson & Do, 2019b), but without agreed-upon formal criteria it is not clear based on what criteria such classifications can be made. We believe that in such cases, ratings can be especially useful for studying iconicity.

Moreover, it is possible to think of ideophones as being gradiently related to other lexical classes (Dingemanse, 2019), with words being more or less ideophonic depending on various properties (phonological markedness, syntactic markedness, semantics, etc.). From this perspective, it is to be expected that highly iconic English words would have features that overlap with ideophones (Dingemanse & Thompson, 2020).

Indeed, the results from our replication study speak to the deep similarities between iconicity in ideophones and iconicity in the general English vocabulary: First, ideophones have been proposed to be structurally marked (Ameka, 2001; Samarin, 1970), and, as we have replicated here using letter frequencies, English words rated high in iconicity are also structurally marked (Dingemanse & Thompson, 2020). Second, ideophones are strongly tied to the senses (Diffloth, 1972; Nuckolls, 1995), and, as demonstrated here, so are English words high in iconicity. Third, ideophones have been linked to word learning in children (e.g., Yoshida, 2012), and so are highly iconic English words (Perry et al., 2015). Fourth and finally, ideophones have been found to be associated with informal discourse (H. Kim et al., 2021; Klamer, 2002; Samarin, 1970), and similarly, we have found iconicity ratings to be correlated with playfulness ratings (Dingemanse & Thompson, 2020). These findings, showing the continuity between iconic words in English and marked iconic words in other spoken languages, demonstrate how research using iconicity rating

datasets such as the one collected here can learn from research on ideophones and vice versa.

6.3. Limitations and recommendations for use

Iconicity ratings are an important part of the methodological toolkit of iconicity research (Motamedi et al., 2019), but it is important to recognize their limitations (Winter & Perlman, 2021a). Most importantly, iconicity ratings underspecify the particular form-meaning links that lead to a rater's intuitions. That is, iconicity ratings just tell us the degree to which people think there is some correspondence between form and meaning, but they do not give clues to the nature of this correspondence. This is not necessarily a problem when performing studies in which claims are not predicated on specific form-meaning pairings, but on iconicity as a general semiotic property of larger clusters of words as a whole. However, it does mean that research with iconicity ratings should be complemented with studies that flesh out the exact nature of form-meaning links. An example of rating studies and statistical studies of the lexicon working in tandem is the finding that English touch words were rated to be high in iconicity (Winter et al., 2017), which subsequently led to the discovery of specific phonemes that are associated with specific textural properties among touch adjectives (Winter, Sóskuthy, et al., 2021).

It is also important to consider whether speaker judgments may be contaminated by factors other than resemblance (Dingemanse & Thompson, 2020; A. L. Thompson et al., 2020). In fact, the ratings themselves provide clear evidence that raters sometimes find it hard to suppress such extraneous factors that have nothing to do with iconicity, at least as we have defined it. For example, *sleepwalk* and *heartburn* received unexpectedly high iconicity ratings in the current study, presumably because of the high semantic transparency of compounds, which may increase the subjective feeling that the form of a word fits its referent. This observation was already made by Dingemanse and Thompson (2020), which led them to perform separate analyses of monomorphemic and multimorphemic words. The examples of *sleepwalk* and *heartburn* clearly show that iconicity ratings are a noisy measure, and one should be careful not to overinterpret the ratings of individual words. This is also why research on iconicity ratings is most reliable when focused on correlations across hundreds or thousands of words, which can help counteract the noisiness inherent in this measure.

Ultimately, all methodological approaches to studying iconicity have their strengths and weaknesses, each warranting the use of complementary methodologies as much as possible. As we have argued here, it is important to recognize that different methods tap into different aspects of the phenomenon of iconicity (Dingemanse et al., 2020; Motamedi et al., 2019). The rating dataset made available here is ideal for correlational studies that allow making generalizations about the lexicon and for explorations into the subjective nature of iconicity.

Acknowledgments

Bodo Winter was supported by the UKRI Future Leaders Fellowship MR/T040505/1. Mark Dingemanse was supported by NWO grant 016.vidi.185.205.

Open Practices Statement

All data and analysis code is made available in the following Open Science Framework repository: https://osf.io/qvw6u/

References

Ahlner, F., & Zlatev, J. (2010). Cross-modal iconicity: A cognitive semiotic approach to sound symbolism. *Sign Systems Studies*, *38*(1/4), 298–348.

Akita, K. (2009). *A grammar of sound-symbolic words in Japanese: Theoretical approaches to iconic and lexical properties of mimetics* [PhD Thesis]. Kobe University.

Akita, K., & Pardeshi, P. (2019). *Ideophones, mimetics and expressives*. John Benjamins.

Ameka, F. K. (2001). Ideophones and the Nature of the Adjective Word Class in Ewe. In F. K. E. Voeltz & C. Kilian-Hatz (Eds.), *Ideophones* (pp. 25–48). John Benjamins.

Amsel, B. D., Urbach, T. P., & Kutas, M. (2012). Perceptual and motor attribute ratings for 559 object concepts. *Behavior Research Methods*, *44*(4), 1028–1041. https://doi.org/10.3758/s13428-012-0215-z

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Bankieris, K., & Simner, J. (2014). Sound symbolism in synesthesia: Evidence from a lexical–gustatory synesthete. *Neurocase*, *20*(6), 640–651.

Bellugi, U., & Klima, E. S. (1975). Aspects of sign language and its structure. In J. Kavanagh & J. Cutting (Eds.), *The role of speech in language* (pp. 171–203). MIT Press.

Bellugi, U., & Klima, E. S. (1976). Two faces of sign: Iconic and abstract. *Annals of the New York Academy of Sciences*, *280*, 514–538. https://doi.org/10.1111/j.1749-6632.1976.tb25514.x

Bergen, B. (2004). The psychological reality of phonaesthemes. *Language*, *80*(2), 290–311.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818–10823. https://doi.org/10.1073/pnas.1605782113

Bolinger, D. (1940). Word affinities. *American Speech*, *15*(1), 62–73. https://doi.org/10.2307/452731

Bolinger, D. (1950). Rime, assonance, and morpheme analysis. *Word*, *6*(2), 117–136. https://doi.org/10.1080/00437956.1950.11659374

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. https://doi.org/10.1037/a0020934

Bredin, H. (1996). Onomatopoeia as a figure and a linguistic principle. *New Literary History*, *27*(3), 555–569.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound

matches, but different shape–taste matches to Westerners. *Cognition*, *126*(2), 165–172.
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.
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.
Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.
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.
Childs, G. T. (1994). African ideophones. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound symbolism* (pp. 178–206). Cambridge University Press.
Clark, H. H., & Gerrig, R. J. (1990). Quotations as demonstrations. *Language*, 764–805.
Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, *125*(3), 452–465.
Croft, W., & van Lier, E. (2012). Language universals without universal categories. *Theoretical Linguistics*, *38*(1–2). https://doi.org/10.1515/tl-2012-0002
Cuxac, C. (1993). Iconicité des langues des signes. *Faits de Langues*, *1*(1), 47–56.
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., … Perlman, M. (2021). Novel vocalizations are understood across cultures. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-89445-4
De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y
Diffloth, G. (1972). Notes on expressive meaning. *Chicago Linguistic Society*, *8*, 440–447.
Diffloth, G. (1994). I: Big, a: Small. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound Symbolism* (pp. 107–114). Cambridge University Press.
Dingemanse, M. (2019). "Ideophone" as a comparative concept. In K. Akita & P. Pardeshi (Eds.), *Ideophones, Mimetics, Expressives* (pp. 13–33). John Benjamins. https://doi.org/10.1075/ill.16.02din
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.
Dingemanse, M., Perlman, M., & Perniss, P. (2020). Construals of iconicity: Experimental approaches to form–meaning resemblances in language. *Language and Cognition*, *12*(1), 1–14. https://doi.org/10.1017/langcog.2019.48

Dingemanse, M., & Thompson, B. (2020). Playful iconicity: Structural markedness underlies the relation between funniness and iconicity. *Language and Cognition*, 1–22. https://doi.org/10.1017/langcog.2019.49

Dryer, M. S. (1997). Are grammatical relations universal? In J. Bybee, J. Haiman, & S. A. Thompson (Eds.), *Essays on Language Function and Language Type* (pp. 115–143). John Benjamins.

Elsen, H. (2017). The two meanings of sound symbolism. *Open Linguistics*, *3*(1), 491–499. https://doi.org/10.1515/opli-2017-0024

Emmorey, K. (2014). Iconicity as structure mapping. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130301. https://doi.org/10.1098/rstb.2013.0301

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods*, *50*(3), 1116–1124.

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: Gesture beats vocalization hands down. *Frontiers in Psychology*, *5*.

Ferrara, L., & Hodge, G. (2018). Language as description, indication, and depiction. *Frontiers in Psychology*, *9*.

Firth, J. R. (1935). The use and distribution of certain English sounds. *English Studies*, *17*(1–6), 8–18.

Flaksman, M. (2017). Iconic treadmill hypothesis. In M. Bauer, A. Zirker, O. Fischer, & C. Ljungberg (Eds.), *Dimensions of Iconicity. Iconicity in Language and Literature* (Vol. 15, pp. 15–38). John Benjamins.

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. Sage publications.

French, P. L. (1977). Toward an explanation of phonetic symbolism. *Word*, *28*(3), 305–322. https://doi.org/10.1080/00437956.1977.11435647

Gasser, M. (2004). The origins of arbitrariness in language. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 434–439). Erlbaum.

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Green, K., & Perlman, M. (2022). Iconic words may be common in early child interactions because they are more engaging. In A. Ravignani, R. Asano, D. Valente, F. Ferretti, S. Hartmann, M. Hayashi, Y. Jadoul, & M. Martins (Eds.), *Proceedings of the Joint Conference on Language Evolution* (pp. 248–255).

Grote, K. (2013). *"Modality relativity": The influence of sign language and spoken language on conceptual categorization* [PhD Thesis]. Hochschulbibliothek der Rheinisch-Westfälischen Technischen Hochschule Aachen.

Hassemer, J., & Winter, B. (2018). Decoding gestural iconicity. *Cognitive Science*, *42*(8), 3034–3049.

Haynie, H., Bowern, C., & LaPalombara, H. (2014). Sound symbolism in the languages of Australia. *PLoS ONE*, *9*(4). https://doi.org/10.1371/journal.pone.0092852

Heise, D. R. (1966). Sound-Meaning Correlations Among 1,000 English Words. *Language and Speech*, *9*(1), 14–27. https://doi.org/10.1177/002383096600900102

Hinojosa, J. A., Haro, J., Magallares, S., Duñabeitia, J. A., & Ferré, P. (2020). Iconicity ratings for 10,995 Spanish words and their relationship with psycholinguistic variables. *Behavior Research Methods*, 1–14. https://doi.org/10.3758/s13428-020-01496-z

Hrushovski, B. (1980). The meaning of sound patterns in poetry: An interaction theory. *Poetics Today*, *2*(1a), 39–56. https://doi.org/10.2307/1772351

Huang, Y.-H., Pratoomraj, S., & Johnson, R. C. (1969). Universal magnitude symbolism. *Journal of Verbal Learning & Verbal Behavior*, *8*(1), 155–156. https://doi.org/10.1016/S0022-5371(69)80028-9

Hutchins, S. S. (1998). *The psychological reality, variability, and compositionality of English phonesthemes* [PhD Thesis]. Emory University.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Phil. Trans. R. Soc. B*, *369*(1651), 20130298.

Jakobson, R., & Waugh, L. R. (1979). *The Sound Shape of Language*. Indiana University Press.

Johansson, N., Anikin, A., & Aseyev, N. (2020). Color sound symbolism in natural languages. *Language and Cognition*, *12*(1), 56–83. https://doi.org/10.1017/langcog.2019.35

Johansson, N., Anikin, A., Carling, G., & Holmer, A. (2019). The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguistic Typology*, *24*(2), 253–310.

Johansson, N., Carr, J. W., & Kirby, S. (2021). Cultural evolution leads to vocal iconicity in an experimental iterated learning task. *Journal of Language Evolution*, *6*(1), 1–25. https://doi.org/10.1093/jole/lzab001

Johansson, N., & Zlatev, J. (2013). Motivations for sound symbolism in spatial deixis: A typological study of 101 languages. *Public Journal of Semiotics*, *5*(1), 3–20.

Joo, I. (2020). Phonosemantic biases found in Leipzig-Jakarta lists of 66 languages. *Linguistic Typology*, *24*(1), 1–12. https://doi.org/10.1515/lingty-2019-0030

Juhasz, B. J., & Yap, M. J. (2013). Sensory experience ratings for over 5,000 mono-and disyllabic words. *Behavior Research Methods*, *45*(1), 160–168.

Käsmann, H. (1992). Das englische Phonästhem sl-. *Anglia-Zeitschrift Für Englische Philologie*, *1992*(110), 307–346.

Kim, H., Winter, B., & Brown, L. (2021). Beyond politeness markers: Multiple morphological and lexical differences index deferential meanings in Korean. *Journal of Pragmatics*, *182*, 203–220. https://doi.org/10.1016/j.pragma.2021.06.006

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214–233. https://doi.org/10.1177/0894439317752406

Kiss, T., Pelletier, F. J., Husic, H., Simunic, R. N., & Poppek, J. M. (2016). A sense-based lexicon of count and mass expressions: The Bochum English Countability Lexicon. *LREC*.

Klamer, M. (2002). Semantically motivated lexical patterns: A study of Dutch and Kambera expressives. *Language*, 258–286.

Knoeferle, K., Li, J., Maggioni, E., & Spence, C. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-05965-y

Köhler, W. (1929). *Gestalt psychology*. Liveright.

Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1693–1710. https://doi.org/10.1080/17470218.2014.989865

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

Kwon, N. (2017). Empirically observed iconicity levels of English phonaesthemes. *Public Journal of Semiotics*, *7*(2), 73–93.

Kwon, N., & Round, E. R. (2015). Phonaesthemes in morphological theory. *Morphology*, *25*(1), 1–27. https://doi.org/10.1007/s11525-014-9250-z

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Lieberth, A. K., & Gamble, M. E. B. (1991). The role of iconicity in sign language learning by hearing adults. *Journal of Communication Disorders*, *24*(2), 89–99. https://doi.org/10.1016/0021-9924(91)90013-9

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01246

Löhr, G. (2021). What are abstract concepts? On lexical ambiguity and concreteness ratings. *Review of Philosophy and Psychology*, *13*, 1–18. https://doi.org/10.1007/s13164-021-00542-9

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.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods*, *45*(2), 516–526.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 1–21. 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

Marchand, H. (1959). Phonetic symbolism in English word-formation. *Indogermanische Forschungen*, *64*, 146–168.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology*, *87*(1/2), 173–188.

Marks, L. E. (1982). Bright sneezes and dark coughs, loud sunlight and soft moonlight. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 177. https://doi.org/10.1037//0096-1523.8.2.177

Marks, L. E. (1989). On cross-modal similarity: The perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 586. https://doi.org/10.1037/0096-1523.15.3.586

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

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC press.

Medler, D. A., Arnoldussen, A., Binder, J. R., & Seidenberg, M. S. (2005). *The Wisconsin perceptual attribute ratings database*. Retrieved from http://www.neuro.mcw.edu/ratings/.

Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1152. https://doi.org/10.1037/a0027747

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Phil. Trans. R. Soc. B*, *369*(1651), 20130299.

Motamedi, Y., Little, H., Nielsen, A., & Sulik, J. (2019). The iconicity toolbox: Empirical approaches to measuring iconicity. *Language and Cognition*, *11*(2), 188–207.

Newman, S. S. (1933). Further experiments in phonetic symbolism. *The American Journal of Psychology*, *45*(1), 53–75.

Nielsen, A., & Dingemanse, M. (2020). Iconicity in word learning and beyond: A critical review. *Language and Speech*, 1–21. https://doi.org/10.1177/0023830920914339

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104. https://doi.org/10.1016/j.cognition.2018.08.014

Nuckolls, J. B. (1995). Quechua texts of perception. *Semiotica*, *103*(1/2), 145–169.

Nuckolls, J. B. (2000). Spoken in the spirit of gesture: Translating sound symbolism in a Pastaza Quechua Narrative. In J. Sherzer & K. Sammons (Eds.), *Translating Native Latin American Verbal Art* (pp. 233–251). Smithsonian Press.

Nuckolls, J. B. (2020). "How do you even know what ideophones mean?": Gestures' contributions to ideophone semantics in Quichua. *Gesture*, *19*(2–3), 161–195. https://doi.org/10.1075/gest.20005.nuc

O'Boyle, M. W., & Tarte, R. D. (1980). Implications for phonetic symbolism: The relationship between pure tones and geometric figures. *Journal of Psycholinguistic Research*, *9*(6), 535–544. https://doi.org/10.1007/BF01068115

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

Orr, J. (1944). On some sound values in English. *British Journal of Psychology*, *35*(1), 1.

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

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

Perlman, M., Little, H., Thompson, B., & Thompson, R. L. (2018). Iconicity in signed and spoken vocabulary: A comparison between American Sign Language, British Sign Language, English, and Spanish. *Frontiers in Psychology*, *9*, 1433. https://doi.org/10.3389/fpsyg.2018.01433

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*.

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*, e12572. https://doi.org/10.1111/desc.12572

Pizzuto, E., & Volterra, V. (2000). Iconicity and transparency in sign languages: A cross-linguistic cross-cultural view. In K. Emmorey & H. Lane (Eds.), *Signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima* (pp. 261–286). Lawrence Erlbaum.

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, *50*(3), 1198–1216.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia–a window into perception, thought and language. *Journal of Consciousness Studies*, *8*(12), 3–34.

Revelle, W. (2021). *psych: Procedures for psychological, psychometric, and personality research*. https://CRAN.R-project.org/package=psych

Rhodes, R. (1994). Aural images. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound symbolism* (pp. 276–292). Cambridge University Press.

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, *43*, 304–307.

Samarin, W. J. (1970). Inventory and choice in expressive language. *Word*, *26*(2), 153–169. https://doi.org/10.1080/00437956.1970.11435590

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

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sidhu, D. M., & Pexman, P. M. (2018a). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*, *25*(5), 1619–1643.

Sidhu, D. M., & Pexman, P. M. (2018b). Lonely sensational icons: Semantic neighbourhood density, sensory experience and iconicity. *Language, Cognition and Neuroscience*, *33*(1), 25–31.

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

Sidhu, D. M., Westbury, C., Hollis, G., & Pexman, P. M. (2021). Sound symbolism shapes the English language: The maluma/takete effect in English nouns. *Psychonomic Bulletin & Review*, 1–9. https://doi.org/10.3758/s13423-021-01883-3

Sidhu, D. M., Williamson, J., Slavova, V., & Pexman, P. M. (2022). An investigation of iconic language development in four datasets. *Journal of Child Language*, *49*(2), 382–396. https://doi.org/10.1017/S0305000921000040

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Stadtlander, L. M., & Murdoch, L. D. (2000). Frequency of occurrence and rankings for touch-related adjectives. *Behavior Research Methods, Instruments, & Computers*, *32*(4), 579–587.

Strik Lievers, F., Bolognesi, M., & Winter, B. (2021). The linguistic dimensions of concrete and abstract concepts: Lexical category, morphological structure, countability, and etymology. *Cognitive Linguistics*. https://doi.org/10.1515/cog-2021-0007

Tarte, R. D. (1974). Phonetic symbolism in adult native speakers of Czech. *Language and Speech*, *17*(1), 87–94. https://doi.org/10.1177/002383097401700109

Taylor, I. K., & Taylor, M. M. (1965). Another look at phonetic symbolism. *Psychological Bulletin*, *64*(6), 413–427.

Thompson, A. L., Akita, K., & Do, Y. (2020). Iconicity ratings across the Japanese lexicon: A comparative study with English. *Linguistics Vanguard*.

Thompson, A. L., & Do, Y. (2019). Unconventional spoken iconicity follows a conventional structure: Evidence from demonstrations. *Speech Communication*, *113*, 36–46. https://doi.org/10.1016/j.specom.2019.08.002

Thompson, B., Perlman, M., Lupyan, G., Sehyr, Z. S., & Emmorey, K. (2020). A data-driven approach to the semantics of iconicity in American Sign Language and English. *Language and Cognition*, *12*(1), 182–202. https://doi.org/10.1017/langcog.2019.52

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.

Verhoef, T., Kirby, S., & de Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive Science*, *40*(8), 1969–1994. https://doi.org/10.1111/cogs.12326

Vinson, D., 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.

Vinson, D., Jones, M., Sidhu, D. M., Lau-Zhu, A., Santiago, J., & Vigliocco, G. (2021). Iconicity emerges and is maintained in spoken language. *Journal of Experimental Psychology: General*.

Vinson, D., Thompson, R. L., Skinner, R., & Vigliocco, G. (2015). A faster path between meaning and form? Iconicity facilitates sign recognition and production in British Sign Language. *Journal of Memory and Language*, *82*, 56–85.

Voeltz, F. E., & Kilian-Hatz, C. (2001). *Ideophones*. John Benjamins Publishing.

Warriner, A. B., & Kuperman, V. (2015). Affective biases in English are bi-dimensional. *Cognition and Emotion*, *29*(7), 1147–1167.

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.

Waugh, L. R. (1993). Against arbitrariness: Imitation and Motivation revived, with consequences for textual meaning. *Diacritics*, *23*(2), 71–87. https://doi.org/10.2307/465317

Waugh, L. R. (1994). Degrees of iconicity in the lexicon. *Journal of Pragmatics*, *22*(1), 55–70. https://doi.org/10.1016/0378-2166(94)90056-6

Werner, H., & Kaplan, B. (1963). *Symbol formation*. John Wiley & Sons.

Wichmann, S., Holman, E. W., & Brown, C. H. (2010). Sound symbolism in basic vocabulary. *Entropy*, *12*(4), 844–858. https://doi.org/10.3390/e12040844

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilcox, S. (2004). Cognitive iconicity: Conceptual spaces, meaning, and gesture in signed language. *Cognitive Linguistics*, *15*(2), 119–147.

Winter, B. (2022). Abstract concepts and emotion: Cross-linguistic evidence and arguments against affective embodiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1870), 20210368. https://doi.org/10.1098/rstb.2021.0368

Winter, B., Oh, G. E., Hübscher, I., Idemaru, K., Brown, L., Prieto, P., & Grawunder, S. (2021). Rethinking the frequency code: A meta-analytic review of the role of acoustic body size in communicative phenomena. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1840), 20200400. https://doi.org/10.1098/rstb.2020.0400

Winter, B., Pérez-Sobrino, P., & Brown, L. (2019). The sound of soft alcohol: Crossmodal associations between interjections and liquor. *PLoS ONE*, *14*(8). https://doi.org/10.1371/journal.pone.0220449

Winter, B., & Perlman, M. (2021a). 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. (2021b). Size sound symbolism in the English lexicon. *Glossa: A Journal of General Linguistics*, *6*(1), Article 1. https://doi.org/10.5334/gjgl.1646

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic? Iconicity in English sensory words. *Interaction Studies*, *18*(3), 433–454. https://doi.org/10.1075/is.18.3.07win

Winter, B., Sóskuthy, M., Perlman, M., & Dingemanse, M. (2021). Trilled /r/ is associated with roughness, linking sound and touch across spoken languages. *Scientific Reports*, *12*, 1035. https://doi.org/10.1038/s41598-021-04311-7

Yoshida, H. (2004). *Iconicity in language learning: The role of mimetics in word learning tasks* [PhD thesis]. Indiana University.

Yoshida, H. (2012). A cross-linguistic study of sound symbolism in children's verb learning. *Journal of Cognition and Development*, *13*(2), 232–265. https://doi.org/10.1080/15248372.2011.573515

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453

Appendix A: Instructions and Sample Trial

In the following, "---" indicates that Participants had to press Next to go on to the next instruction screen.

In this task you will be rating some English words on their "iconicity". Please read the following instructions very carefully as they are important for doing this task.

Some English words sound like what they mean. These words are iconic. You might be able to guess the meaning of such a word even if you did not know English.

Some words that people have rated high in iconicity are "screech", "twirl", and "ooze" because they sound very much like what they mean.

Some words that people have rated moderate in iconicity are "porcupine," "glowing," and "steep," because they sound somewhat like what they mean.

Some words rated low in iconicity are "menu," "amateur," and "are," because they do not sound at all like what they mean.

In this task, you are going to rate words for how iconic they are. You will rate each word on a scale from 1 to 7. A rating of 1 indicates that the word is not at all iconic and does not at all sound like what it means. 7 indicates that the word is high in iconicity and sounds very much like what it means.

It is important that you say the word out loud to yourself, and that you think about its meaning.

If you are unsure of the meaning or the pronunciation of a word, you have the option of skipping it.

Try to focus on the word meaning of the whole word, rather than decomposing it into parts. For example, when rating 'butterfly' think of the insect rather than "butter" and "fly", and rate how well the whole meaning relates to the sound of the whole word "butterfly".

[When you are done with this list of words, you will have the option to do 1-2 additional sets of words, which will earn you bonus pay.]*
*Shown to MTurk participants only.

Please remember to say the word to yourself and to think about the meaning of each word.

Ready to start?
