Two measures are better than one: Combining iconicity ratings and guessing experiments for a more nuanced picture of iconicity in the lexicon

**Abstract**

Iconicity in language is receiving increased attention from many fields, but our understanding of the roles of iconicity in language is only as good as the measures we use to quantify it. We conducted iconicity rating and guessing experiments with 304 Japanese ideophones and prosaic words with sensory meanings, e.g. *fuwafuwa* ‘fluffy’, *jawarakai* ‘soft’. For both word groups, ratings and guesses were positively correlated—suggesting the two measures pick up on the same associations. Ideophones were consistently associated with higher iconicity ratings, but not higher guessing accuracy. We suggest that the structural markedness of ideophones enhances their perceived iconicity in the rating task, but does not provide any advantage (over and above form-meaning associations) in the guessing task. Thus, guesses and ratings could be used together to tease apart the relative contribution of structural markedness to iconic effects. Some ideophones were also poorly guessed, highlighting that construals of iconicity can be language-specific. Finally, we present some methodological contributions, including a new guessing paradigm that improves on the robustness, sensitivity and discriminability of previous approaches, and a reproducible workflow for creating rating and guessing experiments with a Python package, icotools, which we hope will improve comparability between future studies.

**Keywords:** iconicity, ideophones, lexical norms, iconicity ratings, guessability

# 1 Introduction

What does it mean when we say that something is a “good word for” a concept? Certain words, like *klutz*, *schmooze*, and *smush*, seem to “suit” their meanings. We may have a sense that the French *belle* is a more appropriate word for BEAUTY than the Swedish *vacker*. Such intuitions are likely to be somewhat subjective, and influenced by our own linguistic experiences, but we all have them. The idea that certain forms can be particularly suited to certain meanings is known as *iconicity*.

Iconicity–defined here as a perceived resemblance between aspects of form and meaning–is now increasingly recognised as a *design feature of language* (see e.g. Perniss, Thompson, and Vigliocco 2010; Perniss and Vigliocco 2014; Dingemanse et al. 2015 for a review), raising important questions about where iconic mappings come from, and what they do. To answer these questions, we first need to be able to *identify* iconicity in language and, ideally, to *measure* it.

Approaches to measuring iconicity can be divided into descriptive, data-driven, and behavioural methods (Motamedi et al. 2019). Of these, only descriptive approaches target *iconicity* directly, with other methods simply identifying form-meaning association *biases*–whether these be based on resemblance (as in iconicity), co-occurrence (indexicality), statistically driven associations (systematicity), functional advantages, or other factors.

With descriptive approaches, iconicity as a theoretical concept is coded for directly (e.g. Pietrandrea (2002); Hwang et al. (2017); Östling, Börstell, and Courtaux (2018); A. L. Thompson and Do (2019); Voronin (2006); Flaksman (2020)). This is more common and straightforward for sign languages than for spoken languages (A. L. Thompson and Do 2019). It is usually done manually, but Östling et al. -Östling, Börstell, and Courtaux (2018) have shown for sign languages that (for some mappings at least) this can also be automated. Such studies usually involve cross-linguistic data, and aim to uncover universalities in how different languages map sound to meaning–that is, *where* iconicity comes from.

A wider variety of approaches target iconicity *indirectly* through related phenomena. For example, data-driven approaches (e.g. Wichmann, Holman, and Brown (2010); Blasi et al. (2016); Joo (2019); Johansson et al. (2020); Winter et al. (2021)) use comparisons across unrelated languages to identify form-meaning correspondences that occur at rates higher than chance. This requires large amounts of parallel data from distinct phyla, and existing studies usually focus on basic vocabulary–where such data is most readily available. Like descriptive approaches, data-driven approaches identify *specific* form-meaning associations involved in iconic effects, but with the added advantage of not needing to rely on the intuitions and manual labour of individual researchers. However, they *do* rely on the existence of translation equivalents in multiple languages for the words under study. This can pose a particular challenge for highly iconic words, as these often also have highly specific meanings (Lupyan and Winter (2018); Akita (2012)). Thus, these approaches are less likely to be helpful for those with *specific* words for which iconicity measures are required.

Another body of work which capitalises on big data involves the study of *affective iconicity* (e.g. Ullrich et al. 2016; Aryani et al. 2018). These studies use large datasets of affective norms to identify phonological segments loaded with affective meaning (by taking average affectiveness values across all words containing those segments, and identifying segments with values deviating from the global neutral mean). Although limited to associations involving affect, this approach has been successfully used to investigate iconic effects in language processing [Schmidtke and Conrad (2018); Ullrich in prep; Aryani and Jacobs (2018)], and has even been used to explain the affective meaning of poems (Ullrich et al. 2017).

Behavioural approaches involve the collection of data from naive experimental participants, which can reveal something about the iconicity of given form-meaning pairings. The only requirement is the human participants for the study, and the largest amount of iconicity measurements from the widest variety of languages have been collected using these methods. There are two main approaches: iconicity rating tasks and guessing experiments.

In iconicity rating tasks, the concept of iconicity is explained to participants, and participants are asked to rate the iconicity of given form-meaning pairs directly (e.g. Vinson et al. 2008; Perry, Perlman, and Lupyan 2015; Winter et al. 2017; Occhino et al. 2017; A. L. Thompson, Akita, and Do 2020; Punselie 2020). Measurements from these studies have been found to correlate in meaningful ways with factors related to iconicity–e.g. encoding of sensory information (Winter et al. 2017), semantic domains (Punselie 2020), semantic neighbourhood density (Sidhu and Pexman 2018), word class (Perry, Perlman, and Lupyan 2015; A. L. Thompson, Akita, and Do 2020), age of acquisition (Vinson et al. 2008; Perry, Perlman, and Lupyan 2015), facilitated processing (R. L. Thompson, Vinson, and Vigliocco 2009; Occhino, Anible, and Morford 2020), funniness and structural markedness (Dingemanse and Thompson 2020). However, ratings given are influenced by the linguistic experience of participants (Occhino et al. (2017); Sevcikova Sehyr and Emmorey (2019)), so some care is needed to ensure measures are taken from appropriate participant groups.

In guessing experiments, participants are asked to guess the meaning of foreign/novel signs from a choice of two or more alternatives (e.g. Köhler 1929; Tsuru and Fries 1933; Brown, Black, and Horowitz 1955; Kunihira 1971; Ramachandran and Hubbard 2001; Dingemanse et al. 2016; Tzeng, Nygaard, and Namy 2017; Perlman, Dale, and Lupyan 2015; Perlman and Lupyan 2018; Lockwood, Dingemanse, and Hagoort 2016). Signs whose meanings are correctly guessed at rates higher than chance are considered iconic. The guessability of signs has also been linked to facilitated processing (Ormel et al. 2009) and learnability (Perlman and Lupyan 2018; Lockwood, Dingemanse, and Hagoort 2016), and correlated with iconicity ratings (Punselie 2020). However, some signs that were given low iconicity ratings were actually highly guessable, while there were also a few signs that, despite receiving high ratings, were not easily guessed (Punselie 2020: 23). This suggests that guessing and rating tasks may tap into slightly different constructs, or be sensitive to experimental design constraints.

While results from behavioural experiments often require some untangling, we believe that they are particularly valuable sources of iconicity measures for four reasons: (1) they can be applied to any kind of dataset, (2) they are highly scalable, being particularly amenable to online, crowd-sourced data collection, (3) they are tweakable, meaning that we can experiment with different experimental designs to figure out exactly what these measures are telling us, and (4) they are amenable to standardisation, making them a good option for reproducibility. In fact, they are the *only* measures of iconicity that we know of to combine all these things. Moreover, that the measures correlate with each other and with a variety of other meaningful phenomena speaks to their validity.

In this study, we seek to further our understanding of these measures and how they can be useful by comparing and contrasting iconicity ratings and guesses for 304 Japanese words from the sensory vocabulary of Japanese. Sensory language was chosen as this is a domain high in iconicity, and where iconicity offers clear advantages (Winter et al. 2017). The data contained a mix of ideophones–depictive, eye-catching words like *fuwafuwa* ‘soft’ and *pikapika* ‘sparkly’ that “invite and afford the construal of iconic mappings between form and meaning” (Dingemanse 2019: 18)–along with prosaic lexical items, like *yawai* ‘soft’ and *hikaru* ‘shine’, which are not necessarily expected to be lacking in iconicity–especially given their sensory meanings–but which are less obvious about it. We were particularly interested in how the iconicity measures for these words would compare between the rating studies–where participants are asked to make a conscious decision about the iconicity of the words–and the guessing studies–which tap into more unconscious biases. We find that when used together, the two methods shed light on each other, as well as raising pertinent questions about where iconicity effects come from. Our conclusion is that, for many of the questions that we have about iconicity in language, two measures may be better than one.

To discover how to get the most out of each measure, we compare and contrast different rating and guessing paradigms, and in particular introduce a new guessing paradigm that improves on the robustness, discriminability and sensitivity of previous approaches. Finally, to promote the use of these new measures, we introduce a reproducible workflow for automatically generating rating and guessing experiments in a standardised format, via a python package icotools, with support for a variety of stimulus formats (audio, video, and images). This will make the collection of behavioural based iconicity measures quicker, easier, and more comparable between future studies.

# 2 Methods

The data from the study, as well as the code used to produce the experiments and analyses can be found on the Open Science Framework, at <https://osf.io/j57uc/?view_only=d59ffde4bc42467fafb71a904bef8d14>. There is also a html version of this document where you can view the analysis code in line with the text.

## 2.1 Stimuli

The stimuli were 304 Japanese words from the domain of sensory perception, including both ideophones and prosaic lexical items. For all tasks, the words were presented to participants as audio files. The audio files were synthesised using the Google Cloud Text-to-Speech API [(https://cloud.google.com/text-to-speech)](https://cloud.google.com/text-to-speech), with the female Japanese Wavenet voice ja-JP-Wavenet-B. They were then edited using Praat (Boersma and Weenink 2020) to have a flat pitch of 200Hz. Expressive prosody can enhance performance in guessing tasks (Kunihira 1971; Dingemanse et al. 2016), and so using a flat intonation was seen as the best way to ensure comparisons between the words were fair and not influenced by the prosody in a particular recording. We also considered synthesising all the words to have a marked intonation. However, different word lengths would make changes in pitch more dramatic in shorter words compared to longer words, affecting words of different lengths unequally. While flat, synthetic voices may be the best option for comparability, they are not a natural reflection of real-world experiences with iconicity–which involve not only segments but also prosody and even multimodal cues (e.g. accompanying gesture). Thus, if anything the results in this and similar studies likely *underestimate* iconicity in language.

In all studies, participants were instructed to use headphones throughout the experiment, and could not procede to the experiment without first passing a listening test designed to require the use of headphones Woods et al. (2017). The response buttons in each trial also did not appear until *after* the audio files had finished playing. This ensured that participants could not respond without first listening to the audio files.

## 2.2 Data collection

All studies were conducted online, with participants recruited through the crowd sourcing service Prolific (www.prolific.co). The participants were monolingual English speakers residing in England, with no prior knowledge of Japanese. Although rating tasks are often conducted with native speakers, we used English speakers in this study for two reasons: first, to make the results as comparable as possible between the guessing and rating tasks; and second, because our data contained a mix of Standard Japanese and non-standard Japanese words, and we were concerned that using Standard Japanese speakers would lead to a bias towards standard words (Occhino et al. 2017).

The words were split across 11 different experiments, in such a way that each experiment had a balance of words that were expected to be iconic, versus words that were not expected to be iconic. In addition, two ‘practice words’ and three ‘control words’ were included in every experiment. The practice items were used to give participants time to familiarise themselves with the experiment design, and were not included in the results. The control items were chosen from words found to be very iconic/highly guessable in previous studies, and were included as a check to identify participants who were perhaps not performing the task as intended. The first two trials in every experiment were the two ‘practice words’, while the order of the rest of the words in the experiment, including the control words, was randomly varied between participants. After finishing the experiment, participants were asked to describe the task they were performing during the experiment. This was used as an attention check, and to identify participants who had misunderstood the experiment instructions. Strange or lacking task descriptions, particularly when coupled with poor performance on the control items, were used to identify and exclude results from participants who were likely not performing the task as intended.

For the pilot version of the guessing task, in which participants had to match a given form to its correct meaning, data from 3 out of 132 participants was excluded. In one case, the participant indicated that they were choosing the translations based on whether they sounded similar to the Japanese word (rather than based on their meaning). In the other two cases, the participants did not provide a task description and performed poorly on the control items, so the data was excluded as it was not clear that they were paying attention/understood the task correctly. In the new version of the guessing task, in which participants had to match a given meaning to its correct form, data from 3 out of 332 participants was excluded. In all three cases, the participants performed poorly on the control items and either did not provide a task description, or their task descriptions were irrelevant (they commented that the task was ‘interesting’ or ‘difficult’ instead of describing what they were doing in the task).

For the rating task, data from 5 out of 225 participants was excluded. For the ratings, as well as looking at task descriptions and performance on the control items, we calculated the person-total correlation for each participant Curran (2016). The person-total correlation is the correlation between an individual participant’s ratings, and the mean ratings from every other participant. It provides a measure of how well a participant’s responses agree with those of other participants in the task. The criteria for exclusion was two or more of either (a) strange or lacking task descriptions, (b) very low ratings for control items, or (c) low person-total correlations. Criteria (a) was used for 4 exclusions. In two cases, the participants indicated that they were rating the words according to how much they sounded like their English translations, rather than how well they resembled their meanings. Criteria (b) was used in 3 exclusions, and criteria (c) was used in 4 exclusions.

Each experiment took a median of 6 minutes to complete, and participants were paid £0.75 for their participation (based on Prolific’s ‘good’ hourly rate of £7.50). Prolific’s prescreening features were used to ensure that participants who had already participated in one experiment (regardless of whether it was a rating or guessing task) could not participate in any subsequent experiments. Excluding the practice and control items (which every participant guessed and rated), in the final dataset we had a median of 40 guesses per word (range: 15-129) in the pilot guessing experiment; a median of 30 guesses (range: 29-60) per word for the second guessing experiment; and a median of 15 ratings (range: 14-58) per word for the rating task (Motamedi et al. (2019) suggest a benchmark of 10 ratings per word is sufficient).

## 2.3 Pilot guessing experiments: guessing between translations

We piloted some guessing experiments following previous studies (e.g. Dingemanse et al. 2016), in which participants were presented with a Japanese word and asked to guess its meaning from a choice of two possible translations. Some of the ideophones in the data had very specific meanings, which were difficult to capture in a single English word. To provide more context, the translations were given in sentence frames, with the translation word in capitals. The instructions were as follows:

We are interested in how well people can guess the meanings of words in foreign languages. In this experiment, we are interested in your intuitions about the meanings of JAPANESE words. You will hear 25 Japanese words. After listening to each word, try to guess its meaning from a choice of two possible English translations. For context, meanings are given in a sentence, with the intended meaning in CAPITALS. For example, if you think the word could mean BRIGHT in the sentence ‘The sun is BRIGHT’, choose that sentence. If you think it could mean ROUND in the sentence, ‘The sun is ROUND’, choose that sentence. Trust your gut, and good luck!

A sample trial is shown in Figure 2.1.

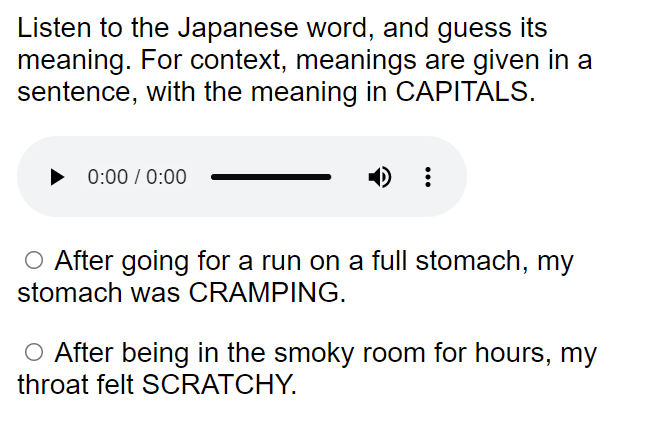


Figure 2.1: Guessing task - guessing between translations

The same formulaic structure, “After X, Y”, was used for all the sentences–where Y is the target perception and X is the event that causes it. As some sensory modalities (e.g. sound) may better lend themselves to the construal of iconicity than others (e.g. taste), correct translations were always paired with foil translations from the same modality (e.g. sound translations were only paired with other sound translations, and taste translations with other taste translations, etc.). To make the difficulty of the task comparable between different trials, only words that were neither synonyms nor antonyms of the correct translation were chosen as foils. Finally, the length of the sentences was also kept consistent (to a difference of no more than 5 characters) between pairings of translations and foils. To test the robustness of the guessing procedure to different choices of translation and foil words, the same words were tested multiple times with different translations and foils. We were concerned that participants could choose particular translations because of how they sound (particularly if they sounded similar to the Japanese word), rather than based on their meanings as we intended. So, to minimise any potential effects from this, as much as possible phonologically distinct synonyms were chosen as alternative translations (e.g. SMALL|TINY, BIG|LARGE).

Pilot results showed that in several cases the guessability of a word did differ considerably depending on the translation or foil word used. Some examples are shown in Table 2.1.

**Table 2.1: Effect of different foils and translations on guessing results for hakkiri ‘CLEAR HEADED’ and piiN ‘LONG’**

| **Answer sentence** | **Foil sentence** | **correct** | **incorrect** |
| --- | --- | --- | --- |
| After a nice relaxing bath, my head felt completely CLEAR. | After missing breakfast and lunch, I was completely STARVING. | 3 | 12 |
| After a nice relaxing bath, my mind felt completely CLEAR. | After being outside all day long, I was incredibly COLD. | 14 | 6 |
| After pulling it and pulling it, the rubber band was very LONG. | After cutting up the tomatoes, my hands were RED. | 15 | 5 |
| After pulling it and pulling it, the rubber band was very STRETCHED | After cutting up the tomatoes, my hands were RED. | 4 | 16 |

We hypothesised that, as well as considering the meaning of the translations and foils, participants may be driven to one option over another because of the way the words sound (particularly if they sounded similar to the Japanese word being tested), or because of lexical features of the words like frequency, valence, arousal, or any other associations they may have with these words.

While we could not do anything about the sound of the words (other than using multiple translations), we realised that we could eliminate at least some of the complications caused by other properties of the words if we flipped the design of the guessing experiments on its head, so that instead of choosing the correct English meaning for a given Japanese word, participants chose the correct Japanese word for a given English meaning. In this way, there was only one English word involved in the task, and the options participants chose between were simply two Japanese words which they knew nothing about, other than how they sound. This design was also more comparable to the rating task–where participants are also given the meaning of the word first. We therefore decided to use this new design for the rest of the guessing experiments.

## 2.4 Guessing experiments: guessing between words

For the new version of the guessing experiments, we minimised the amount of English words involved in the task by having participants match a single English translation to the corresponding word in Japanese, choosing between two possible Japanese words. The instructions used were as follows:

We are interested in how well people can guess words in foreign languages. In this experiment, you will be asked to match the English translation with the corresponding word in JAPANESE, guessing from a choice of two Japanese words.

A sample trial is shown in Figure 2.2.

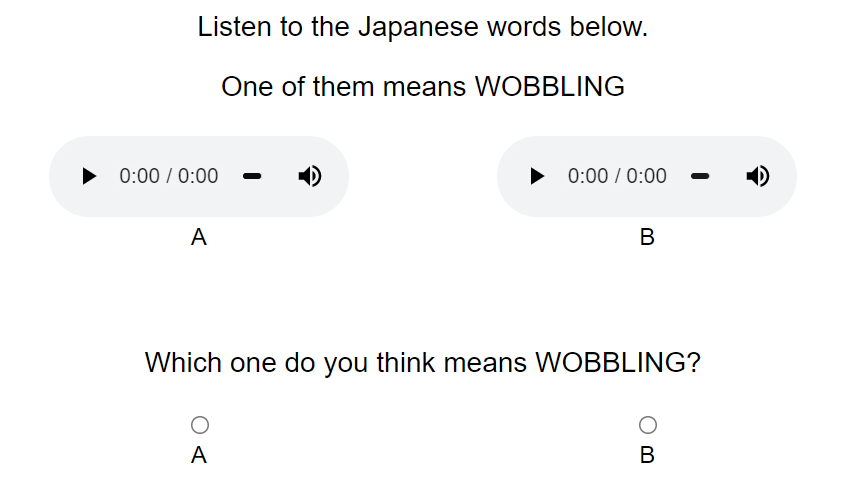


Figure 2.2: Guessing task - guessing between words

This time, instead of using sentence frames for context, we kept the amount of English to a minimum by using only single-word translations where possible, and where more context was needed this was provided in brackets, as in Figure 2.3.

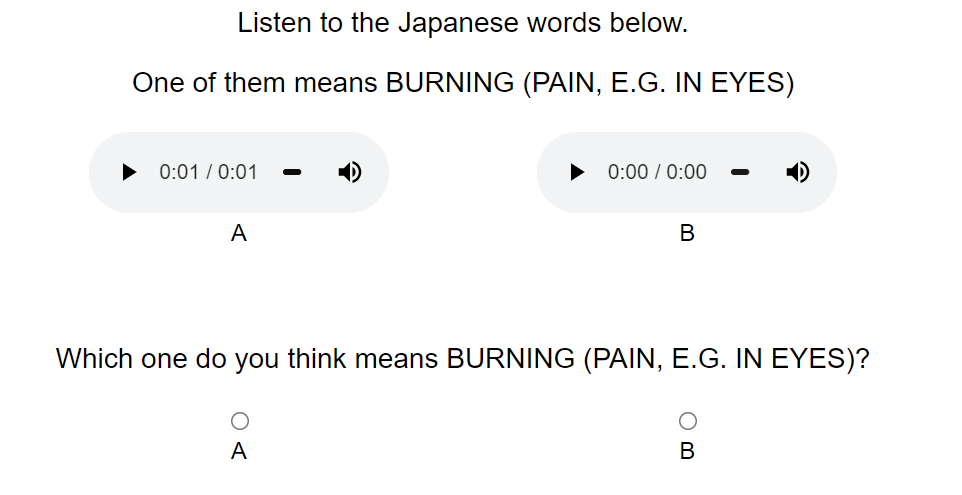


Figure 2.3: Guessing task - guessing between words

The incorrect ‘foil’ word in each item was a nonsense Japanese word artificially generated to be as phonologically distinct as possible from the correct word, while still conforming to Japanese phonology. For every word, three different foil words were generated by substituting each consonant with its top three most phonologically distant consonants in the Japanese sound system. Phonological distances were calculated using the feature matrix provided in PHOIBLE (Moran and McCloy 2019), with the distance between two sounds being equal to the sum of the distances between each of their feature values. Distances were calculated as follows:

The distance between two feature values that are identical is 0. The distance between two opposing values (+/- or -/+) is 1—except if the feature involved is length or voicing, then the distance is 0.5, The distance between two feature values when one of them is 0 (= not applicable) is 0.25.

Since /a/ is often the most phonologically distant vowel from many of the vowels in the Japanese sound system, to avoid the foil words all having similar vowels, a different approach was used for substituting vowels. Instead of using feature values, vowels were substituted with the vowel obtained by rotating the vowel space 180 clockwise or counterclockwise. So /u/ is substituted with either /a/ or /i/, and /a/ is substituted with either /i/ or /u/ (see Figure 2.4). Finally, reduplicated words were ‘unreduplicated’ when creating foil words. For example, the foils for *fuwafuwa* were *watʃi*, *nati* and *ridːu* (not *watʃiwatʃi*, *natinati* and *ridːuridːu*).

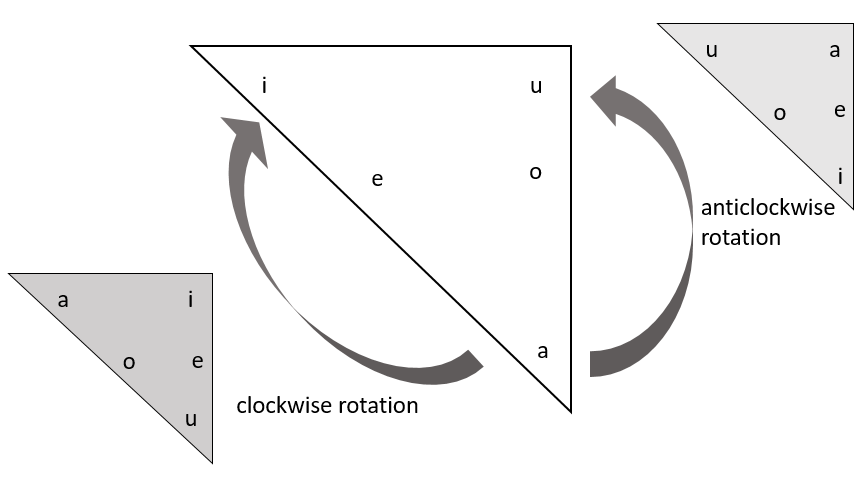


Figure 2.4: Vowel substitutions for the creation of foil words

By using foil words that sound as different as possible to the target word, we hoped to improve the sensitivity of the measure to any iconicity in the target word. Our logic was that if the target word *is* iconic for a concept, then a word that sounds very different to it should be a bad fit for that same concept–making the choice between the two easier. To test whether this was really the case, we ran one of the experiments a second time using foils that were randomly chosen from among the other trial items in the experiment, rather than phonologically distinct foils.

## 2.5 Rating task

We also conducted an iconicity rating task with the same words. In the rating task, the concept of iconicity was defined to participants as “when a word and its meaning resemble one another”, using the English examples *wiggle*, *jiggle*, and *wriggle*. The instructions stated that “Even people who do not speak any English can get a sense of the meaning of these words”, and contrasted them with words like *walk* and *run* whose meanings are “not so intuitive”. The full instructions given were as follows:

Some words seem to ‘fit’ their meanings. For example, consider the English words wiggle, jiggle, and wriggle.  
We have an intuitive sense of the meanings of these words, because there is a resemblance between the words and their meanings.  
Even people who do not speak any English can get a sense of the meaning of these words.  
Words like walk and run on the other hand are not so intuitive; people who do not know any English would not be able to guess what these words mean.  
In this task, you will listen to some Japanese words, and we will tell you their meanings. You will then be asked to judge whether there is a resemblance between the word and its meaning.

Participants had to listen to the Japanese words and were told their meanings, then asked to rate the resemblance between the word and its meaning on a scale from 0 ‘No resemblance’ to 6 ‘Strong resemblance’ (see Figure 2.5). Previous studies (Perry, Perlman, and Lupyan 2015) also used a negative end of the scale (corresponding to a ‘bad resemblance’). However, follow-up analyses found that the negative end of the scale was both underused and inconsistently used, suggesting that anti-iconic relationships are difficult to assess (Motamedi et al. 2019). For this reason, we decided to only use a positive scale in this study.

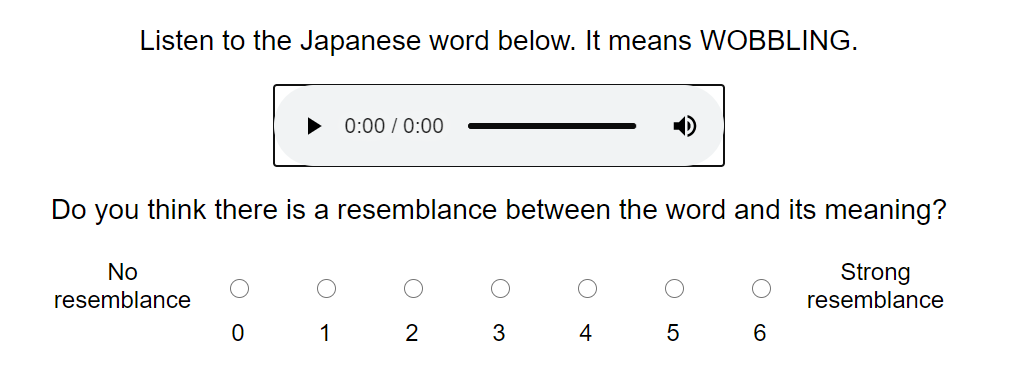


Figure 2.5: Rating task

As with the guessing experiments, the translation shown for each participant was varied randomly between a set of (wherever possible) phonologically distinct synonyms.

## 2.6 Creation of Icotools

To make the collection of guesses and ratings quicker and easier for future studies, while also improving comparability between studies, we created a python package, icotools, with functions that can be used to automatically generate rating and guessing experiments (using the meaning-to-word design from the second guessing experiments). The package has three functions: a rater function, a guesser function, and a foiler function. The rater and guesser functions create rating and guessing experiments as described in Sections 2.4 and 2.5. They take two csv files as input. The first is a wordlist containing the items in the experiment (including trials, practice items, control items, and sets of translations to use for them), while the second file is used for customisation (e.g. to specify how many items to include in each experiment, the location of media files, custom instructions and custom exit questions, etc.).

For each experiment, the functions produce three files–a html file, a php file, and a csv file. These files can then be uploaded to a server, the link to the html files can be shared with participants, and as the participants complete the experiments their responses will be written to the csv file by the php file. The functions also produce a summary csv file with all the details of the experiments (i.e. the trials in each experiment, as well as the translations and, where applicable, foils used). An R script is provided to collate the responses for all experiments into one long tidy dataset after data collection is complete.

The foiler function can be used to generate a list of opposite sounding foil words for each item in a guessing experiment, following the method described in Section 2.4. It is also possible to specify your own foil words, or to choose the foils randomly from among the other items in the experiment. For a full description of how to use icotools, please see the ReadMe at <https://anonymous.4open.science/r/IcoTools-2CEA/> (this is an anonymised link to the github repository: it will later be replaced with the correct link).

# 3 Results

Statistical analyses were conducted using R version 4.0.3 (2020-10-10).

## 3.1 Robustness to different choices of translations and foils

We first investigated whether ratings and guesses were robust to different choices of translations (for ratings and guesses) and foils (for guesses only). For each word separately, we performed Fisher’s Exact Tests on the guessing data, and ANOVAs on the rating data, to determine whether the guessing performance or ratings given differed significantly between different translations and foils. As the ANOVAs are less robust to small amounts of data than the Fisher’s Exact Tests, for the rating data we excluded translations with fewer than 10 ratings from this particular analysis.

In the first pilot of the guessing task–where participants were guessing between English translations–11 out of the 45 words tested (=24%) showed significant differences in their guessability when either the translation and/or the foil was changed. In the second version of the guessing task–where participants were guessing between Japanese words–39 out of 303 words tested (=13%) showed significant differences in their guessability when either the translation and/or the foil was changed. In the rating task, 11 out of 88 words tested (=13%) showed significant differences in their iconicity rating when the translation word was changed.

## 3.2 Sensitivity and discriminability

To test whether using phonologically distinct foils in the guessing experiments improved their sensitivity to iconicity, a subset of 30 words were tested twice–once with phonologically distinct “opposite” foils, and once with random foils. The number of these words which were guessed significantly above chance was higher in the experiment with opposite foils than in the experiment with random foils. Using opposite foils, half of the words in the sample (15/30) were guessed significantly above chance, whereas with the random foils this number fell to just 1 in 5 words (6/30). Only two words were guessed significantly above chance in the random foil condition, but not in the opposite foil condition. However, both were guessed numerically above chance in the opposite foil condition as well (with accuracies of 52% and 57%). The differences between the two conditions are shown in Figure 3.1. A chi-squared test confirmed the differences to be significant (χ2(1)=4.69,p=.03).

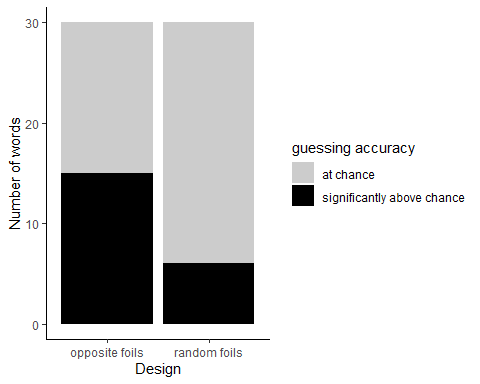


Figure 3.1: Words guessed significantly above chance using opposite versus random foils

We also compared the distributions of guessing accuracies and mean iconicity ratings between the different methods. Since the iconicity ratings in this study were collected from naive participants who did not speak any Japanese, whereas rating studies more commonly use native speaker participants, we added one more plot to the figure showing the distribution of iconicity ratings in a separate study by Thompson et al. (2020). Thompson et al. collected iconicity ratings for Japanese words from native Japanese speaking participants. Their dataset contained a wider variety of words than the current dataset, and they also used a different scale. Their scale went from -5 ‘antiiconic’ to 5 ‘iconic’, with 0 being labelled as ‘arbitrary’. The scale in this study started at 0 (also labelled as ‘arbitrary’, or “no resemblance between the word and its meaning”), and went up to 6 for “strong resemblance between the word and meaning”. To make the data more comparable between the two studies, and because previous studies have found the negative ratings unreliable (Motamedi et al. 2019), I have filtered the data from Thompson et al. to only contain responses using the part of the scale between 0 and 5. I have also filtered the words to only contain verbs, adverbs, and adjectives from the Yamato stratum. The Yamato stratum refers to the native lexicon of Japanese. Thompson et al.’s study included both words from the Yamato stratum, as well as Sino-Japanese and foreign words. This study only included Yamato words, mainly verbs, adverbs and adjectives. Both the data in this study and in Thompson et al. 2020 contains a mix of ideophones and non-ideophones within the Yamato stratum. Finally, ratings from both studies were transformed so that they varied between 0 and 1, to match with the guessing accuracies. The resulting distributions are shown in Figure 3.2

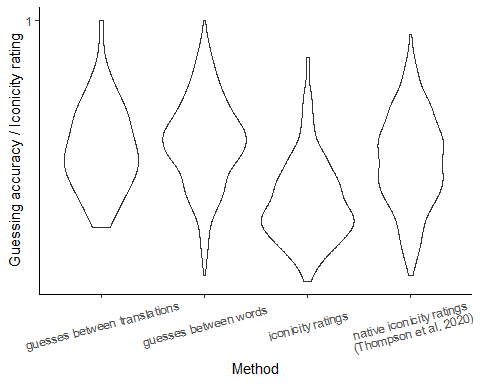


Figure 3.2: Discriminability of different measures

Focusing first on the data from the current study, we can see that a greater spread of measures is provided by the guesses compared to the ratings, and that within the guesses, the method of guessing between words provides a better spread of measures than guessing between translations–both at the very high end of the scale (highly guessable words) and at the very low end of the scale (poorly guessable words). However, if we look at the ratings from native Japanese speakers, these have a pretty equivalent spread to the guesses between words. Their distribution is slightly narrower at the very top end of the scale. However, the data from the current study was chosen from a domain known to be high in iconicity, whereas the data from Thompson et al. was more varied, with proportionally less ideophones. Thus, this could simply be a reflection of the differences between the two datasets rather than relating to the methods used.

## 3.3 Agreement

Figure 3.3 compares the guessing accuracies and mean iconicity ratings, for both ideophones and non-ideophones. The words in the pilot study where participants were guessing between translations were different to the words in the subsequent guessing and rating studies, so here only the results from the later guessing study (where participants guessed between words) are shown. At the top end of the scale, the two measures agree–ideophones score higher than non-ideophones for both measures. However, at the bottom of the scale something strange happens. While for the ratings, the bottom of the scale is dominated by non-ideophones (as expected), for the guesses, the bottom of the scale is actually dominated by *ideophones*. That is, ideophones are simultaneously guessed *better* and *worse* than non-ideophones. We will return to this in the discussion.

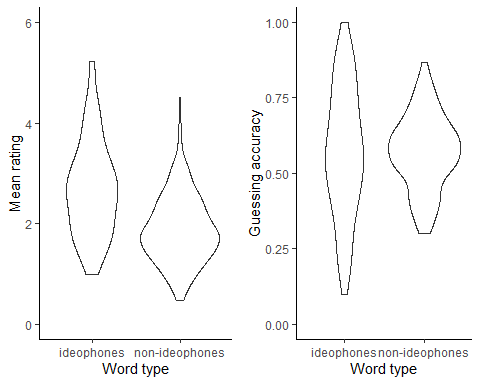


Figure 3.3: Comparison of guesses and ratings, for ideophones and non-ideophones

Setting aside the strange distribution of guesses for the ideophones, it’s notable that both measures suggest that the sensory lexicon of Japanese is for the most part *iconic*, rather than arbitrary. In the guesses, this translates to the majority of the words being guessed slightly *above*, rather than at or below chance. In the ratings, this translates to the words receiving mean ratings higher than zero (the arbitrary point on the scale)–though some of these ‘iconic’ ratings are likely an overstatement, since all of the words received positive iconicity ratings and yet from the guessing results we can see that many of these were not guessed any differently to chance.

To test how well the measures agree on the iconicity of individual words, iconicity ratings and guessability scores for the same words were transformed to z-scores so that they could be directly compared. Figure 3.4 plots matching z-scores for the guesses and ratings against eachother, to show the agreement between the two measures. In the figure, ideophones are represented by blue dots, and non-ideophones by red dots.

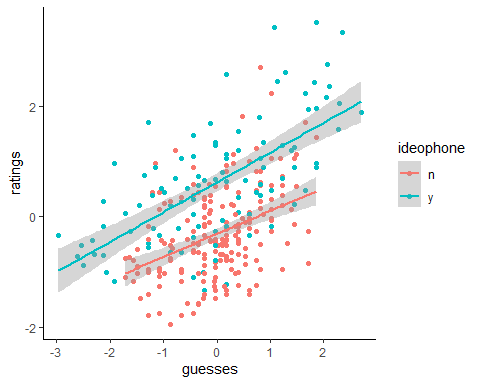


Figure 3.4: Agreement between guesses and ratings

For both ideophones and non-ideophones, there was a strong correlation between the ratings and the guesses. This correspondence was slightly better for the ideophones (r = 0.66, 95% CI [0.44, 0.79], 95% CI [0.53, 0.76], t(99) = 8.68, p < .001) than for the non-ideophones (r = 0.41, 95% CI [0.29, 0.52], t(201) = 6.45, p < .001). To explore this relationship further, we created two linear regression models: one predicting guesses from ratings, ideophone status, and the interaction between these two factors, and a second model predicting ratings from guesses, ideophone status, and the interaction between these two factors. Both models again used the z-scores rather than the raw ratings and guesses. The models were created using the R package lmerTest (version 3.1.3) (Kuznetsova, Brockhoff, and Christensen 2017). The output of the first model (predicting guesses from ratings) is shown in Table 3.1.

**Table 3.1: Linear regression model predicting guesses from ratings, ideophone status, and the interaction between these two factors.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Guesses* | | | |
| *Predictors* | *Estimates* | *std. Error* | *Statistic* | *p* |
| (Intercept) | 0.13 | 0.06 | 2.13 | **0.034** |
| ratings | 0.41 | 0.07 | 5.59 | **<0.001** |
| ideophone [y] | -0.64 | 0.11 | -5.65 | **<0.001** |
| ratings \* ideophone [y] | 0.39 | 0.11 | 3.69 | **<0.001** |
| Observations | 304 | | | |
| R2 / R2 adjusted | 0.325 / 0.318 | | | |

The model confirmed that ratings significantly predict guesses, with higher ratings leading to higher guesses. There was also a main effect of ideophone, and an interaction effect between ideophone and rating. The model is easiest to interpret by plotting the interaction, shown in Figure 3.5.

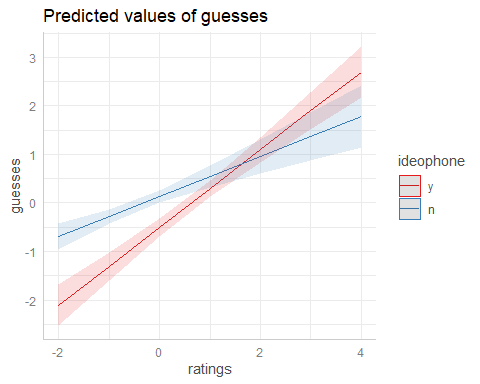


Figure 3.5: Interaction between ratings and ideophone when predicting guesses

The plot shows that higher ratings lead to better guessing accuracy for both ideophones and non-ideophones. The interaction effect appears because in the case of words with *low* iconicity ratings, guessing accuracy for ideophones is relatively poorer than for non-ideophones. At the top of the scale, highly rated ideophones also appear to be guessed slightly better than non-ideophones with comparable ratings. However, the difference here does not appear to be significant as the confidence intervals for the two lines overlap. We will return to this in the discussion.

For comparison, we also created a second model predicting ratings from guesses, shown in Table 3.2.

**Table 3.2: Linear regression model predicting ratings from guesses, ideophone status, and the interaction between these two factors.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Ratings* | | | |
| *Predictors* | *Estimates* | *std. Error* | *Statistic* | *p* |
| (Intercept) | -0.31 | 0.05 | -5.81 | **<0.001** |
| guesses | 0.42 | 0.07 | 6.12 | **<0.001** |
| ideophone [y] | 0.93 | 0.09 | 10.10 | **<0.001** |
| guesses \* ideophone [y] | 0.12 | 0.09 | 1.40 | 0.163 |
| Observations | 304 | | | |
| R2 / R2 adjusted | 0.432 / 0.426 | | | |

The model shows that guesses significantly predict ratings, as does ideophone status, but that there was no interaction between these two predictors. This is shown in Figure 3.6. The lines for ideophones and non-ideophones are parallel, indicating that differences in guesses correspond to differences in ratings *in the same way* for both ideophones and non-ideophones. However, the line for the ideophones is directly *above* the line for the non-ideophones, indicating that ideophones are rated higher in iconicity than non-ideophones–even when guessed at the same accuracies.

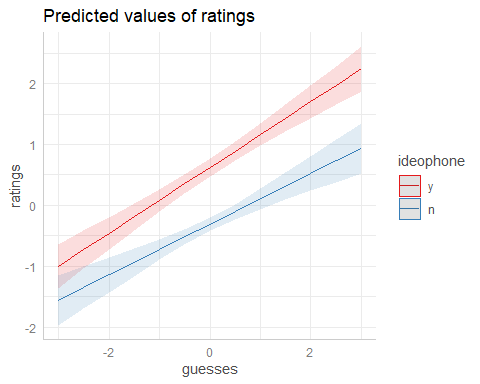


Figure 3.6: Relationship between ideophone status, guessability, and predicted rating

Finally, the lower correlation between guesses and ratings obtained for non-ideophones compared to ideophones suggests that participants may not have been as consistent or reliable when rating non-ideophones compared to when rating ideophones. To further investigate this, we calculated the person-total correlation for the rating data (Curran 2016; Motamedi et al. 2019), for ideophones and non-ideophones separately. Figure 3.7 shows the agreement between raters (expressed as the correlation between each individual rating, and the by-item average) for ideophones and non-ideophones respectively. The dots indicate the mean while the lines indicate 95% confidence intervals. Again, since these ratings were from non-Japanese speaking participants, we have also added data from native Japanese speaking participants from Thompson et al. (2020) as a comparison.

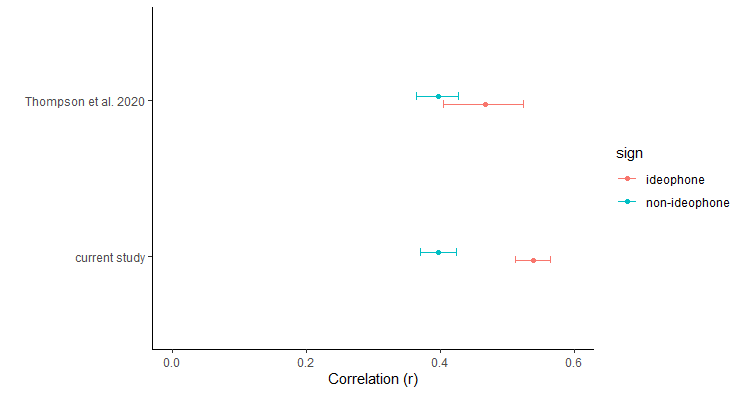


Figure 3.7: Consistency between participants for iconicity ratings of ideophones versus non-ideophones

In both studies, participants appear to be more consistent with eachother when rating ideophones than when rating non-ideophones, though for Thompson et al. the overlap of the confidence intervals indicates that the difference was not significant. There does not appear to be any difference in the consistency of ratings *between* the English and the Japanese speakers, as both the ideophone lines and the non-ideophone lines overlap between the two groups.

# 4 Discussion

## 4.1 Relationship between ratings, guesses, ideophones, and iconicity

For both ideophones and non-ideophones, we found that the guesses and ratings from non-speakers were strongly correlated in the *same way*. This suggests that the two measures are indeed tapping into *something* in common, which we suggest are form-meaning associations. It’s notable that these associations are found in both ideophones *and* non-ideophones (even if they are more common in the former), and that there are also ideophones that participants had no strong associations with–highlighting that there is no straightforward relationship between ideophones and iconicity; variable levels of iconicity can be found in both ideophones and prosaic words (see also Dingemanse et al. 2016; Brown, Black, and Horowitz 1955; Kunihira 1971).

However, ideophones were consistently rated *higher* in iconicity than non-ideophones—even when guessed at the same accuracies. We believe this is due to the majority of ideophones in the dataset being reduplicated. The two lowest rated ideophones–*sappari* and *syippori*–were among the few in the data that do not use a reduplicated template. Previous work has found a relationship between iconiticity ratings and structural markedness (Dingemanse and Thompson 2020). Ordinarily when we use words, the form of the words is not important–only their content. However, when iconicity is involved (and also funniness, as this paper found) the form of the words *is* important, and structural markedness may serve as a metacommunicative cue to signal this. The authors hypothesise that “Structural markedness confers a selective advantage on words intended to be iconic… as their recognisability would make them more fit to survive processes of cultural transmission in which the recognition of such intentions is functionally important” (Dingemanse and Thompson 2020: 218). The results from the rating task in this study suggest that structural markedness (in this case, reduplication) does indeed lead to the perception of iconicity. Interestingly, however, it does not seem to confer any advantage–over and above that provided by associations between form and meaning–in the guessing task. What the ratings seem to be picking up on is both form-meaning associations *and* structural markedness. If they were only picking up on structural markedness, then we would not expect the ratings to vary with the guesses in the same way for both ideophones and non-ideophones. However, if they were only picking up on form-meaning associations (like the guesses seem to), then we would not expect ideophones to consistently receive higher iconicity ratings, even when guessed at the same accuracies.

While ideophones and non-ideophones with high iconicity ratings were generally guessed equally well, there was an interaction effect such that ideophones with low iconicity ratings were guessed *worse* than non-ideophones with comparable ratings. Low iconicity ratings in non-ideophones generally correspond to chance guessing accuracy–reflecting a lack of associations–whereas equivalent ratings in ideophones correspond to *below chance* guessing accuracy–reflecting *negative associations* masked in the rating results by the tendency to inflate ratings for ideophones. The result highlights the subjective nature of form-meaning associations, and that by taking above-chance guessing accuracy as a measure of iconicity we are only picking up on those associations that happen to correspond–and crucially not to conflict–between (in this case) English and Japanese speakers. This is something researchers should be aware of, because it means that decisions about *whom* to collect iconicity measurements from are very important.

Poor guessing accuracies could be explained by conflicting associations. For example, one person commented that *zarazara* as a word for ‘ROUGH’ sounds wrong to them because they associate the word *zara* with the clothing store ZARA, and clothes are soft. These could be masking either a lack of resemblance between form and meaning, or a resemblance between form and meaning that is ignored in the face of the conflicting association. Of course, conversely one could argue that high guessing accuracies could then *also* be explained by non-resemblance based associations. Fortunately, the chances of English speakers having the *same* form-meaning association as Japanese speakers based on something random (like the name of a clothing store) should be much smaller than the chances of them having a *different* association. Thus, if they do exist, we predict that the majority of non-resemblance based associations should be reflected in poorer rather than higher guessing accuracies. It is interesting that conflicting associations appear to have a stronger effect in ideophones (some of which were guessed at accuracies significantly below chance) than in non-ideophones (which were guessed below chance only rarely). Perhaps because participants are more likely to view form-meaning relationships in ideophones as meaningful (due to their structural markedness), they are more likely to attach conflicting associations to ideophones, leading to a stronger sense that the ideophone should correspond to the *opposite* of the given meaning.

Finally, the comparison between guesses and ratings highlights a bias in the design of the rating task, as all the words in the data were in the end rated as “slightly” iconic (even by non-speakers), whereas the guessing results show that a good number of these were not guessed any better than chance. This is probably an artefact of the design of the rating task, in which “arbitrary” is only 1 of the 7 or 10 points on the scale–the rest indicating varying degrees of form-meaning resemblance. This is not so problematic, as the ratings still correlate with guessing accuracy, indicating that the relative differences in iconicity still hold. However, it highlights that interpretations of particularly *low* iconicity ratings as reflections of iconicity should be viewed with caution.

To sum up, by directly comparing iconicity ratings and guessing accuracies for both ideophones and non-ideophones, we are able to better understand how these measures relate to each other and to iconicity. Differences between iconicity ratings of ideophones and non-ideophones–even when guessability is the same–highlight the role of structural markedness in enhancing perceived iconicity. Consistencies between guesses and ratings within these word types reassure us that, despite this, form-meaning association biases play a large role in explaining each measure. The unexpected result of ideophones also being poorly guessed reminds us that these biases are just that–subjective–and thus, while good guessing accuracies can be interpreted as positive evidence of language-external iconicity, average or poor guessing accuracies *cannot* be interpreted as negative evidence of iconicity of either kind (language-internal or external). Similarly, while high ratings are likely to be a good indication of iconicity, researchers should be cautious of taking lower ratings indicating “slight” iconicity at face value.

## 4.2 Towards a synthesis of measures

Given the limitations of each measure discussed above, for a full picture of iconicity in the lexicon, a synthesis of measures is recommended. Iconicity ratings from native speakers can tell us which mappings are meaningful to them, while comparisons with guessability for non-speakers can establish which of these mappings are likely to be meaningful on their own, versus which *become* meaningful through experience with the rest of the lexicon (language-external versus language-internal iconicity). Occhino, Anible, and Morford (2020) found differential effects of these two types of mappings in sign processing by signers of different proficiencies. They found that both types of iconicity were helpful to signers of a lower proficiency, but that for highly proficient signers only the most pervasive, language-external mappings made a difference to processing.

Since structural markedness appears to boost ratings but not guessability, the two measures could also be used together to tease apart the influence of structural markedness versus form-meaning resemblances in driving iconic effects. Iconic effects are generally assumed to be driven by form-meaning resemblances, but since these often go together with structural markedness it would be useful to know the relative contribution of each factor in different situations and tasks. Since structural markedness is also subjective, the choice of participants used in the rating task will be important. Finally, having guesses alongside ratings would allow researchers to make better decisions about the cut-off point between “high” and “low” iconicity ratings, as these could be motivated by correlations with guessing accuracy, rather than e.g. arbitrarily dividing the scale at the half-way point, or using percentiles tied to particular datasets.

Guesses and ratings could also be used to inform results from descriptive or data-driven measures, and vice-versa. For example, the psychological reality and relative *strength* of form-meaning mappings described in lexicons could be tested using guessing or rating tasks with nonwords (e.g. Kwon 2017). Similarly, descriptive (semiotic) analysis, along with cross-linguistic comparisons where possible could be used to understand and explain why some mappings are stronger or more universal than others. Conversely, behavioural measures could be used to tease apart some of the possible explanations for cross-linguistic biases. The two main contenders for explaining these biases are form-meaning associations or communicative pressures (Blasi et al. 2016). Since communicative pressures are unlikely to play a role in tasks where participants are simply judging or choosing words, behavioural measures could be used to distinguish these two hypotheses, thus allowing us to better understand the mechanisms behind results in these cross-linguistic studies.

## 4.3 Methodological improvements and evaluation

We also sought to evaluate different rating and guessing paradigms, and were able to make some methodological improvements–particularly to the guessing design. We contrasted two different guessing paradigms–one where participants are given a word and have to match it to the correct meaning from two translations, and another where participants are given a meaning and have to match it to the correct word from two words. Content-wise, choosing between words is more comparable to the design of a rating task, as both give the meaning first and involve only one English translation. This is important as the use of translations leads to several complications in the experiment design. First, English-speaking participants can be expected to have more associations with English words than with unknown Japanese words, which could make the task more sensitive to the particular translations chosen. In addition, care has to be taken when choosing foils as well, as they should preferably be from the same semantic domain, of a similar length of characters, and not synonyms or antonyms of the correct translation. This makes the guessing between translations design more difficult to use, particularly in small semantic domains (e.g. taste).

In contrast, a task in which participants guess between words is much easier to operationalise. This design can be easily adopted to any semantic domain or any kind of data, and the foils can be chosen either randomly from the other trials or selected based on specific criteria. In our case we chose foils that were phonologically as distinct as possible from the word being tested. The first reason for this was theoretical: in a task designed to pick up on form-meaning resemblance, words that sound very different should not be suited to that meaning, enhancing the effect of iconicity. The second was practical: our stimuli had artificially flat intonations, which has been shown to decrease performance in guessing studies (Kunihira 1971; Dingemanse et al. 2016). By creating opposite sounding words, we tried to increase performance in the guessing task despite the degraded stimuli.

When comparing results from the two designs, choosing between words is again the more favourable of the two. First, as expected changing the translation or foil used led to less inconsistent results when choosing between words compared to when choosing between translations. Second, the range of guessing scores is much wider when choosing between words compared to choosing between translations, with more words guessed at high accuracy levels, and fewer words around chance level. This is especially true when the foils used are maximally phonologically distinct. However, for datasets that are relatively higher in iconicity to begin with–e.g. gestures or hand signs–using random foils will likely already produce good results. The better results in the choosing between words design can be easily explained, as the task was comparatively easier than in the choosing between translations design, where participants were restricted to choosing between two similar options (non-antonyms from the same semantic domain). The options in the choosing between words design are more distinct–even maximally so. In sum, the choosing between words design is both more robust, more sensitive to iconicity, and more discriminating between lesser and greater amounts of iconicity, so this is the design that we recommend and that we have implemented in icotools.

We also compared our iconicity ratings from non-speakers with iconicity ratings from native speakers in Thompson et al. (2020). We found that, although using non-speakers resulted in a lesser spread of measures (suggesting that native speakers are more confident in giving high and low ratings than non-native speakers), when evaluating the agreement between raters, ratings from non-speakers were just as reliable as ratings from native speakers–as well as being fairly robust to different choices of translation. Ratings from non-speakers were more reliable for ideophones than for non-ideophones–but this is not to say that the ratings for non-ideophones were unreliable. Rather, the ratings for ideophones were *particularly* reliable, probably because raters were more likely to have stronger intuitions about these words due to their structural markedness, and because they are more likely to be iconic in the first place. The same pattern was seen in the ratings from native speakers, although it did not reach statistical significance. Since both types of ratings are fairly reliable, our recommendation would be to choose participants for rating tasks according to how the ratings are going to be used. For example, to predict language processing in adult native speakers of a language, it would make sense to collect ratings from other adult native speakers of the same language.

## 4.4 Towards a reproducible workflow

Finally, with icotools we introduce a reproducible workflow for conducting rating and guessing experiments. This will make it easier for researchers to run these tasks, while enhancing comparability between future studies. Researchers can use the same paradigms to confirm current results, ask new questions, and run new analyses that build on existing findings. For example, this study has examined iconicity in the sensory vocabulary of Japanese, and found that the majority of words tested were actually guessed slightly above chance. With icotools, future researchers could run the exact same guessing experiment with different languages and in different semantic domains, to establish whether this is also true in other languages, and in other areas of the lexicon.

# 5 Conclusions

This study has asked how we can make existing measures of iconicity even better, by critically evaluating and comparing two of the most promising behavioural measures of iconicity–iconicity ratings and guessing tasks. In comparing the two measures, we were able to get a better sense of the relative strengths and weaknesses of each. Our main finding is that the two measures are more informative when used together than when used in isolation. In isolation, it is difficult to determine the degree to which high iconicity ratings are driven by form-meaning associations versus structural markedness, and to have an objective sense of which ratings are “low” or “high”. In isolation, guesses provide only a narrow window into iconicity of a particular kind–strong, language-external associations. However, in *combination* ratings and guesses together can provide a picture of iconicity at different levels (both language internal and external), and in different degrees (low and high), while also uncovering how perceptions of iconicity are driven by both form-meaning associations and structural markedness. We have also made some methodological improvements to the guessing paradigm, improving on the robustness, sensitivity and discriminability of previous approaches. Finally, we introduce a reproducible workflow for creating rating and guessing tasks in the form of the Python package icotools, so that future studies will be able to use these measures to build on current findings as well as asking new questions about what iconicity does in language, and how it does it.

# 6 Data Availability Statement

The data from the study, as well as the code used to produce the experiments and analyses can be found on the Open Science Framework, at <https://osf.io/j57uc/?view_only=d59ffde4bc42467fafb71a904bef8d14>. There is also a html version of this document where you can view the analysis code in line with the text.

# References

Akita, Kimi. 2012. “Toward a Frame-Semantic Definition of Sound-Symbolic Words: A Collocational Analysis of Japanese Mimetics.” *Cognitive Linguistics* 23 (1): 67–90.

Aryani, Arash, Markus Conrad, David Schmidtke, and Arthur Jacobs. 2018. “Why ’Piss’ Is Ruder Than ’Pee’? The Role of Sound in Affective Meaning Making.” *PLOS ONE* 13 (6): e0198430.

Aryani, Arash, and Arthur M. Jacobs. 2018. “Affective Congruence Between Sound and Meaning of Words Facilitates Semantic Decision.” *Behavioral Sciences* 8 (6): 56.

Blasi, Damián E., Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. 2016. “Sound–Meaning Association Biases Evidenced Across Thousands of Languages.” *Proceedings of the National Academy of Sciences* 113 (39): 10818–23.

Boersma, Paul, and David Weenink. 2020. “Praat: Doing Phonetics by Computer [Computer Program].” <http://www.praat.org/>.

Brown, Roger W., Abraham H. Black, and Arnold E. Horowitz. 1955. “Phonetic Symbolism in Natural Languages.” *The Journal of Abnormal and Social Psychology* 50 (3): 388.

Curran, Paul G. 2016. “Methods for the Detection of Carelessly Invalid Responses in Survey Data.” *Journal of Experimental Social Psychology* 66: 4–19.

Dingemanse, Mark. 2019. “’Ideophone’ as a Comparative Concept.” In *Ideophones, Mimetics, and Expressives*, edited by Kimi Akita and Prashant Pardeshi, 13–33. Amsterdam: John Benjamins Publishing Company.

Dingemanse, Mark, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen, and Padraic Monaghan. 2015. “Arbitrariness, Iconicity, and Systematicity in Language.” *Trends in Cognitive Sciences* 19 (10): 603–15.

Dingemanse, Mark, Will Schuerman, Eva Reinisch, Sylvia Tufvesson, and Holger Mitterer. 2016. “What Sound Symbolism Can and Cannot Do: Testing the Iconicity of Ideophones from Five Languages.” *Language* 92 (2): e117–33.

Dingemanse, Mark, and Bill Thompson. 2020. “Playful Iconicity: Structural Markedness Underlies the Relation Between Funniness and Iconicity.” *Language and Cognition* 12 (1): 203–24.

Flaksman, Maria A. 2020. “Stanislav Voronin’s Universal Classification of Onomatopoeic Words: A Critical Approach.” *Dyskurs* 6 (4): 131–49.

Hwang, So-One, Nozomi Tomita, Hope Morgan, Rabia Ergin, Deniz İlkbaşaran, Sharon Seegers, Ryan Lepic, and Carol Padden. 2017. “Of the Body and the Hands: Patterned Iconicity for Semantic Categories\*.” *Language and Cognition* 9 (4): 573–602.

Johansson, Niklas Erben, Andrey Anikin, Gerd Carling, and Arthur Holmer. 2020. “The Typology of Sound Symbolism: Defining Macro-Concepts via Their Semantic and Phonetic Features.” *Linguistic Typology* -1 (ahead-of-print).

Joo, Ian. 2019. “Phonosemantic Biases Found in Leipzig-Jakarta Lists of 66 Languages.” *Linguistic Typology* 0 (0).

Köhler, Wolfgang. 1929. *Gestalt Psychology*. New York: Liveright.

Kunihira, Shirou. 1971. “Effects of the Expressive Voice on Phonetic Symbolism.” *Journal of Verbal Learning and Verbal Behavior* 10 (4): 427–29.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. “lmerTest Package: Tests in Linear Mixed Effects Models.” *Journal of Statistical Software* 82 (13): 1–26.

Kwon, Nahyun. 2017. “Empirically Observed Iconicity Levels of English Phonaesthemes.” *Public Journal of Semiotics* 7 (2): 73–93.

Lockwood, Gwilym, Mark Dingemanse, and Peter Hagoort. 2016. “Sound-Symbolism Boosts Novel Word Learning.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42 (8): 1274.

Lupyan, Gary, and Bodo Winter. 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.

Moran, Steven, and Daniel McCloy, eds. 2019. *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. <https://phoible.org/>.

Motamedi, Yasamin, Hannah Little, Alan Nielsen, and Justin Sulik. 2019. “The Iconicity Toolbox: Empirical Approaches to Measuring Iconicity.” *Language and Cognition* 11 (2): 188–207.

Occhino, Corrine, Benjamin Anible, and Jill P. Morford. 2020. “The Role of Iconicity, Construal, and Proficiency in the Online Processing of Handshape.” *Language and Cognition* 12 (1): 114–37.

Occhino, Corrine, Benjamin Anible, Erin Wilkinson, and Jill P. Morford. 2017. “Iconicity Is in the Eye of the Beholder.” *Gesture* 16 (1): 99–125.

Ormel, Ellen, Daan Hermans, Harry Knoors, and Ludo Verhoeven. 2009. “The Role of Sign Phonology and Iconicity During Sign Processing: The Case of Deaf Children.” *Journal of Deaf Studies and Deaf Education* 14 (4): 436–48.

Östling, Robert, Carl Börstell, and Servane Courtaux. 2018. “Visual Iconicity Across Sign Languages: Large-Scale Automated Video Analysis of Iconic Articulators and Locations.” *Frontiers in Psychology* 9.

Perlman, Marcus, Rick Dale, and Gary Lupyan. 2015. “Iconicity Can Ground the Creation of Vocal Symbols.” *Royal Society Open Science* 2 (8): 150152.

Perlman, Marcus, and Gary Lupyan. 2018. “People Can Create Iconic Vocalizations to Communicate Various Meanings to Naïve Listeners.” *Scientific Reports* 8 (1): 1–14.

Perniss, Pamela, Robin Thompson, and Gabriella Vigliocco. 2010. “Iconicity as a General Property of Language: Evidence from Spoken and Signed Languages.” *Frontiers in Psychology* 1 (227): 1–15.

Perniss, Pamela, and Gabriella Vigliocco. 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.

Perry, Lynn K., Marcus Perlman, and Gary Lupyan. 2015. “Iconicity in English and Spanish and Its Relation to Lexical Category and Age of Acquisition.” *PloS One* 10 (9): e0137147.

Pietrandrea, Paola. 2002. “Iconicity and Arbitrariness in Italian Sign Language.” *Sign Language Studies*, 296–321.

Punselie, Stella. 2020. “Iconicity in Ideophones.” Bachelor's thesis, Nijmegen: Radboud University.

Ramachandran, Vilayanur S., and Edward M. Hubbard. 2001. “Synaesthesia–a Window into Perception, Thought and Language.” *Journal of Consciousness Studies* 8 (12): 3–34.

Schmidtke, David, and Markus Conrad. 2018. “Effects of Affective Phonological Iconicity in Online Language Processing: Evidence from a Letter Search Task.” *Journal of Experimental Psychology: General* 147 (10): 1544.

Sevcikova Sehyr, Zed, and Karen Emmorey. 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): 208–34.

Sidhu, David M., and Penny M. Pexman. 2018. “Lonely Sensational Icons: Semantic Neighbourhood Density, Sensory Experience and Iconicity.” *Language, Cognition and Neuroscience* 33 (1): 25–31.

Thompson, Arthur Lewis, Kimi Akita, and Youngah Do. 2020. “Iconicity Ratings Across the Japanese Lexicon: A Comparative Study with English.” *Linguistics Vanguard* 0 (0).

Thompson, Arthur Lewis, and Youngah Do. 2019. “Defining Iconicity: An Articulation-Based Methodology for Explaining the Phonological Structure of Ideophones.” *Glossa: A Journal of General Linguistics* 4 (1).

Thompson, Robin L., David P. Vinson, and Gabriella Vigliocco. 2009. “The Link Between Form and Meaning in American Sign Language: Lexical Processing Effects.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35 (2): 550.

Tsuru, Shigeto, and H. Fries. 1933. “A Problem in Meaning.” *Journal of General Psychology* 8: 281–84.

Tzeng, Christina Y., Lynne C. Nygaard, and Laura L. Namy. 2017. “The Specificity of Sound Symbolic Correspondences in Spoken Language.” *Cognitive Science* 41 (8): 2191–2220.

Ullrich, Susann, Arash Aryani, Maria Kraxenberger, Arthur M. Jacobs, and Markus Conrad. 2017. “On the Relation Between the General Affective Meaning and the Basic Sublexical, Lexical, and Inter-Lexical Features of Poetic Texts—A Case Study Using 57 Poems of H. M. Enzensberger.” *Frontiers in Psychology* 0.

Ullrich, Susann, Sonja A. Kotz, David S. Schmidtke, Arash Aryani, and Markus Conrad. 2016. “Phonological Iconicity Electrifies: An ERP Study on Affective Sound-to-Meaning Correspondences in German.” *Frontiers in Psychology* 0.

Vinson, David P., Kearsy Cormier, Tanya Denmark, Adam Schembri, and Gabriella Vigliocco. 2008. “The British Sign Language (BSL) Norms for Age of Acquisition, Familiarity, and Iconicity.” *Behavior Research Methods* 40 (4): 1079–87.

Voronin, Stanislav V. 2006. *Osnovy Fonosemantiki [The Fundamentals of Phonosemantics]*. Moscow: Moskva Lenand.

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown. 2010. “Sound Symbolism in Basic Vocabulary.” *Entropy* 12 (4): 844–58.

Winter, Bodo, Marcus Perlman, Lynn K. Perry, and Gary Lupyan. 2017. “Which Words Are Most Iconic?” *Interaction Studies* 18 (3): 443–64.

Winter, Bodo, Márton Sóskuthy, Marcus Perlman, and Mark Dingemanse. 2021. “Trilled/r/Is Associated with Textural Roughness in the Vocabularies of Spoken Languages (OSF Preprint).” https://doi.org/<https://osf.io/jq2sz/>.

Woods, Kevin J. P., Max Siegel, James Traer, and Josh H. McDermott. 2017. “Headphone Screening to Facilitate Web-Based Auditory Experiments.” *Attention, Perception & Psychophysics* 79 (7): 2064–72.