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

# 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); Thompson and Do (2019); Voronin). This is more common and straightforward for sign languages than for spoken languages (Thompson and Do 2019). It is usually done manually, but Östling et al. (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. (n.d.)) 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 tend to also have highly specific meanings (Lupyan and Winter (2018); Akita (2012)). Thus, such approaches are less likely to be helpful for those with existing data 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; Thompson, Akita, and Do 2020; Punselie 2020). Measurements from these studies are shown to correlate in meaningful ways with factors related to iconicity–e.g. encoding of sensory information (Winter et al. 2017), semantic neighbourhood density (Sidhu and Pexman 2018), age of acquisition (Vinson et al. 2008; Perry, Perlman, and Lupyan 2015), facilitated processing (Thompson, Vinson, and Vigliocco 2009; Occhino, Anible, and Morford 2020), word class (Perry, Perlman, and Lupyan 2015; Thompson, Akita, and Do 2020), semantic domain (Punselie 2020), funniness and structural markedness (Dingemanse and Thompson 2020). However, ratings given are strongly influenced by the linguistic experience of participants (Occhino et al. (2017); Sevcikova Sehyr and Emmorey (2019)), which could make comparing measurements between studies difficult.

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.

Despite these complications, behavioural measures of iconicity are particularly appealing 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 better understand what these guesses and ratings are measuring by comparing and contrasting iconicity ratings and guesses for 300 Japanese words from the sensory vocabulary of Japanese. Sensory language was chosen as this is a domain high in iconicity, 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 regular 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 investigate 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 [(https://pypi.org/project/icotools/)](https://pypi.org/project/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 different studies.

# 2 Method

## 2.1 Stimuli

The stimuli were 300 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..[[1]](#footnote-1)

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 done with native speakers, we used English speakers in order to make the results as comparable as possible between the guessing and the rating tasks.

The words were split across 12 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 rating task, no data was excluded. There were some task descriptions where participants claimed to be rating based on whether the Japanese word sounded similar to its English translation, rather than whether they thought it was an appropriate word for that meaning. However, as their actual ratings did not seem to reflect this and were similar to those of other participants, their responses were not excluded. The vast majority of participants’ task descriptions indicated that they had understood the task very well.

For the first version of the guessing task, in which participants had to match a given form to its correct meaning, data from three 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, no data was excluded. Participants’ task descriptions indicated that they had understood the task very well and they also performed well on the control items.

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. For each word, we collected 30 guesses and 30 ratings from different participants.

## 2.3 Guessing experiments – guessing between translations

In the first guessing experiments, 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 used are shown below, and a sample trial is shown in Figure 2.1.

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!

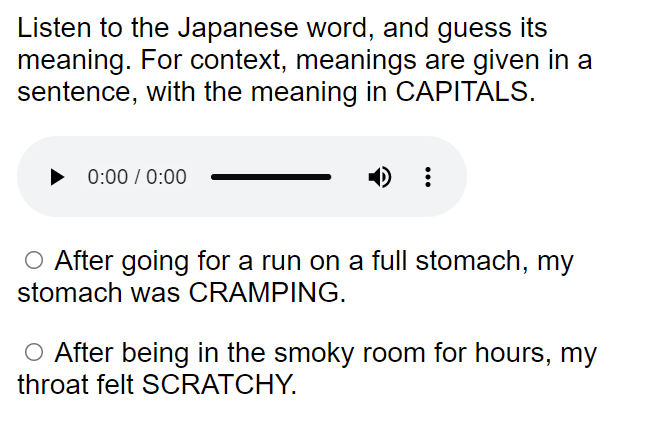


Figure 2.1: Guessing task - guessing between translations

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. Since we wanted participants to choose the translations based on their *meanings*, and not how they *sound* (e.g. we did not want participants choosing a particular option simply because it sounded similar to the Japanese word), as much as possible, phonologically distinct synonyms were chosen as alternative translations (e.g. AFRAID|SCARED, TINY|SMALL).

Pilot results showed that in several cases the guessability of a word differed 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, 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 had to choose 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, and a sample trial is shown in Figure 2.2.

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.

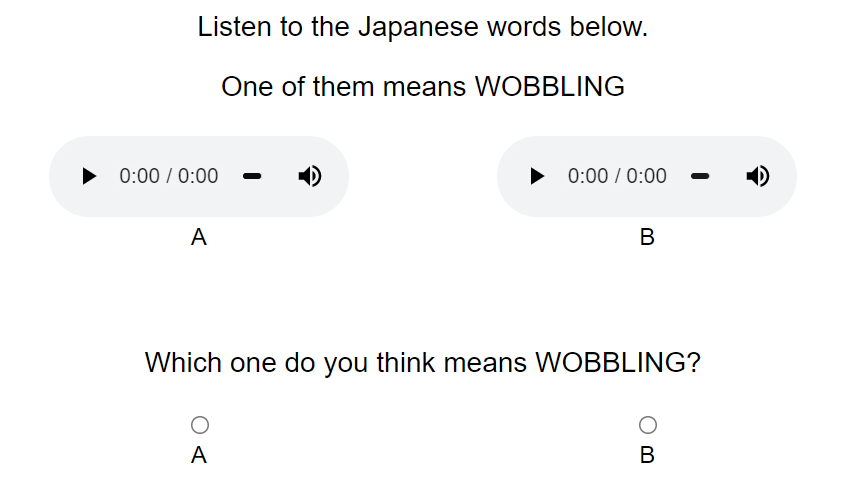


Figure 2.2: Guessing task - guessing between words

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

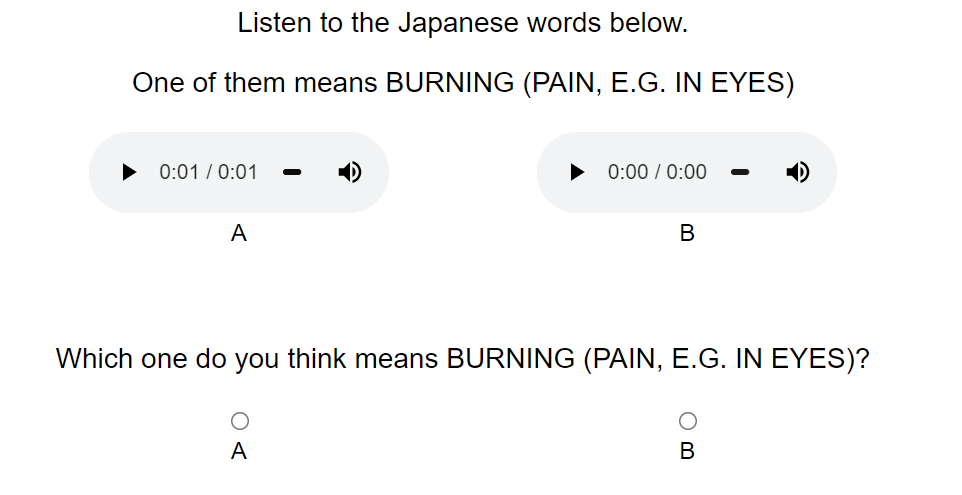


Figure 2.3: Guessing task - guessing between words

The foil words were artificially generated to be as phonologically distinct as possible from the answer 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 Phonological Corpus Tools software (Hall, Mackie, and Lo 2019), with the feature system provided by Mielke (2012). Vowels were substituted by rotating the vowel space 180 degrees, so that /a/ > /u/, /i/ > /a/, /u/ > /i/, /e/ > /o/, and /o/ > /e/. Finally, reduplicated words were ‘unreduplicated’ in their foil words. For example, the foils for *fuwafuwa* were *rizyu*, *gityu* and *nitsu* (not *rizyurizyu*, *gityugityu* and *nitsunitsu*).

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 potential iconicity in the target word, as 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 with foils that were randomly chosen from the other trials in the experiment, instead of using 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’. A sample trial is shown in Figure 2.4.

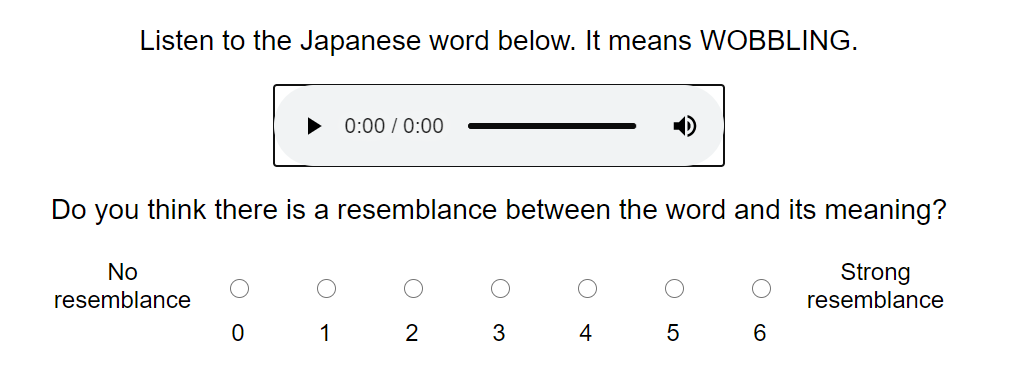


Figure 2.4: Rating task

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

# 3 Results

## 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 guessing performance or ratings given differed significantly between different translations and foils.

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–15 out of the 131 words tested (=11%) showed significant differences in their guessability when either the translation and/or the foil was changed. In the rating task, 12 out of 92 words tested (=13%) showed significant differences in their iconicity rating when the translation word was changed[[2]](#footnote-2).

## 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 (with a sample of 30 guesses per word, an accuracy greater than 2/3 is needed to be sure the true accuracy is above 0.5) 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).

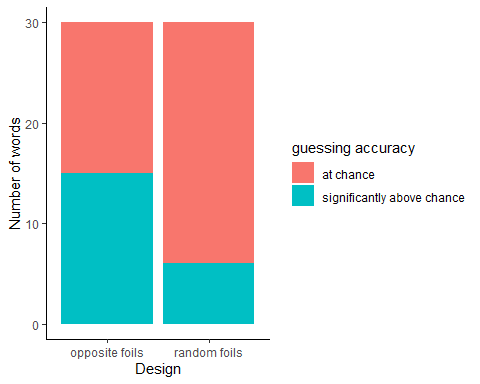


Figure 3.1: Words guessed significantly above chance using opposite versus random foils

We also compared the distributions of guessing accuracies and mean iconicity ratings between the different methods, shown in Figure 3.2. 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, I 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, 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, and predominantly these were 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.

In order to plot the distributions on the same scale, the iconicity ratings were transformed so that they varied between 0 and 1–to match the guessing accuracies. However, note that the locations of the distributions are not comparable between the guessing and rating studies, as in the rating studies a score of 0 corresponds to a lack of iconicity, whereas a lack of iconicity in the guesses would correspond to a score of 0.5. The main purpose of the figure is to compare the *shape* of the distributions between the different measures.

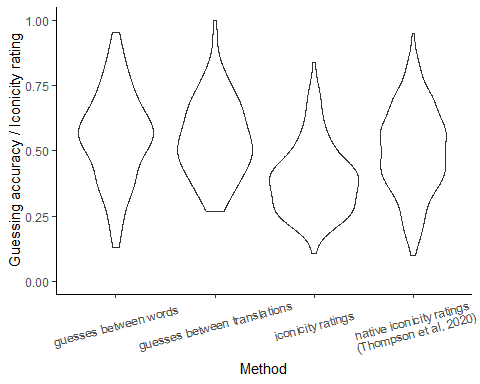


Figure 3.2: Discriminability of different measures

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 from a more varied dataset, with proportionally less ideophones. Thus, this could well be simply 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 in the ratings, the bottom of the scale is dominated by non-ideophones (as expected), in 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.

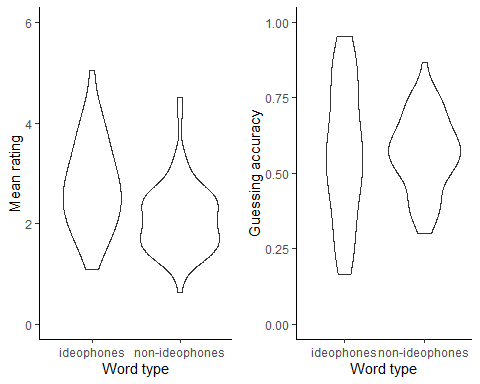


Figure 3.3: Comparison of guesses and ratings

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 ratings, this translates to all the words receiving mean ratings higher than zero (the arbitrary point on the scale). In the guesses, this translates to the majority of the words being guessed slightly *above*, rather than at or below 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.

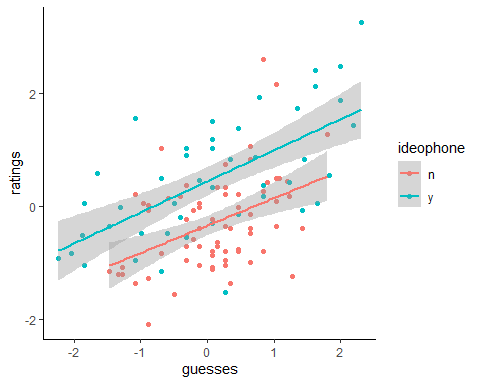


Figure 3.4: Agreement between guesses and ratings

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 = 0.65, 95% CI [0.44, 0.79], t(43) = 5.57, p < .001) than for the non-ideophones (r = 0.43, 95% CI [0.23, 0.60], t(71) = 4.06, 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 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)

0.20

0.10

2.04

0.044

ratings

0.39

0.12

3.38

0.001

ideophone [y]

-0.48

0.16

-2.92

0.004

ratings \* ideophone [y]

0.37

0.16

2.34

0.021

Observations

118

R2 / R2 adjusted

0.338 / 0.320

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.

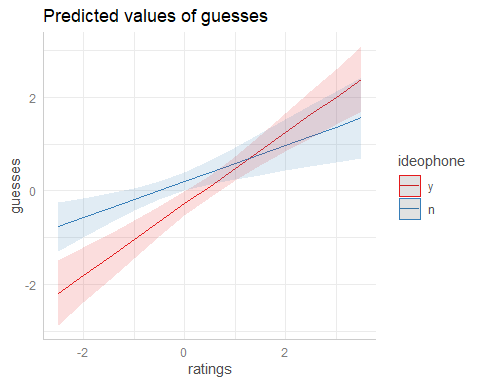


Figure 3.5: Interaction between ratings and ideophone when predicting guesses

The plot shows that higher ratings lead to better guessing accuracy for both ideophones and non-ideophones, although this is much more dramatic for the ideophones compared to the non-ideophones. That is, ideophones with lower ratings are guessed much poorer than non-ideophones with lower ratings, and similarly ideophones with higher ratings are guessed much better than non-ideophones with higher ratings. This could explain why the correlation between the ratings and guesses was stronger for the ideophones than for the non-ideophones; though in both cases ratings and guesses positively correlate.

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)

-0.34

0.09

-3.76

<0.001

guesses

0.48

0.12

3.88

<0.001

ideophone [y]

0.78

0.15

5.28

<0.001

guesses \* ideophone [y]

0.06

0.16

0.42

0.677

Observations

118

R2 / R2 adjusted

0.411 / 0.396

The model shows that guesses significantly predict ratings, as did ideophone status, but 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. We will return to this in the discussion.

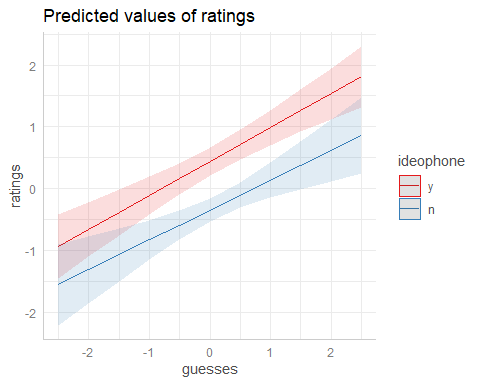


Figure 3.6: Relationship between ideophone status, guessability, and predicted rating

Another possible explanation for the weaker correlation between guesses and ratings obtained for the non-ideophones is 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, I 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

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 association biases. It’s notable that these form-meaning association biases are found in both ideophones *and* non-ideophones, 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.

Ideophones also behaved differently to non-ideophones in the guessing results. In particular, they were more likely to be guessed both very *well*, and very *poorly*. This counterintuitive result raises important issues about the interpretation of guessing accuracy as a reflection of iconicity. In the beginning of the study, we defined ideophones as words that “invite and afford the construal of iconic mappings between form and meaning”. When there is no construal of a relationship between form and meaning, then we expect guesses to be essentially random. When there *is* a construal of a relationship between form and meaning, then we expect guesses to be non-random. That ideophones are guessed both very well and very poorly thus supports the idea that they facilitate the construal process. That this works *both* ways highlights the caveat mentioned in the beginning of this study which is that these methods simply pick up on *associations*, which may or may not be motivated by perceived form-meaning *resemblances* (i.e. iconicity). That these resemblances are *perceived* is also important as it reflects the idea that iconicity is not an objective property of signs, but rather a subjective process involving construal (Occhino et al. 2017; Occhino, Anible, and Morford 2020).

Poor guessing accuracies could thus be explained by form-meaning association biases based on something *other than* resemblance (e.g. 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 true resemblance between form and meaning. It is impossible to know from guessing results alone. Of course, it follows 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.

That the poor guessing results were also correlated with lower ratings suggests that iconicity ratings also likely reflect form-meaning *associations* rather than form-meaning *resemblances* specifically. This is despite the instructions in the task specifically asking about resemblances. That participants in rating studies do not distinguish iconicity from indexicality is hardly surprising, given it’s a distinction even seasoned semioticians can find a struggle. Nor is it necessarily problematic–whether the association is based on resemblance or other factors, the end result it the same, a bias. However, it is something researchers should be aware of, because it means that decisions about *whom* to collect iconicity measurements from are very important, since associations are much more dependent on world experience than resemblances (see also Motamedi et al. 2019: 17).

The notion of construal also complicates the interpretation of guessing accuracies at and below chance-level. Occhino et al. (2020: 117) provide an excellent illustration of this in the ASL sign for dance, which involves an inverted V-handshape. On its own, the meaning of this sign is unlikely to be transparent to non-signers. However, a signer will recognise the inverted-V handshape from a network of signs involving the construal of a pair of legs mapped to a pair of extended fingers–such as in the sign STAND, which is likely to be transparent to non-signers. Thus, iconicity can be both language-internal (as in the sign for DANCE) and/or language-external (as in the sign for STAND). However, guessing studies and rating tasks with non-speakers (as in the current study) will only pick up on examples of language-external iconicity, while language-internal iconicity falls under the radar.

It is also possible that different construals could lead to non-speakers having exactly the *opposite* iconic mapping to the mapping in the language under study, which could also lead to poor guessing accuracy. This is perhaps more likely in spoken than in signed languages, particularly where phoneme inventories are small. The most famous example comes from vowel-size sound symbolism. The most common mapping relates high vowels–like /i/–to smallness, and low vowels–like /a/–to largeness (Ultan 1978). However, a few languages also show the *reverse* mapping–two examples are Korean and Bahnar (Shinohara and Kawahara 2010; Diffloth 1994). Different construals make it possible for both mappings to be perceived as iconic to speakers, and in fact this is likely since the reverse mappings are actually reported in ideophones–which speakers generally perceive as iconic (Thompson, Akita, and Do 2020). For example, a construal of the change from /i/ to /a/ as involving an *expansion* of the space in the oral cavity could lead to the /i/-small, /a/-large mapping. However, one could also construe the same event as involving the *contraction* of the space occupied by the tongue, resulting in an /i/-large, /a/-small mapping. Interestingly, despite the conflicting mapping in their own language, when evaluating non-words Korean speakers still show a preference for the /i/-small, /a/-big mapping (Shinohara and Kawahara 2010). This is encouraging as it suggests that conflicting mappings in the participants own languages are unlikely to hide iconic effects in cases where the naturalness of the mapping is very strong, as is certainly the case in the /i/-small, /a/-big mapping, demonstrated by its overwhelming presence cross-linguistically (Ultan 1978).

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) ratings as reflections of iconicity should be taken with a grain of salt.

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. Guessing and rating comparisons could also be useful to explore the influence of structural markedness on iconicity effects. Since structural markedness appears to boost ratings but not guessability, the two measures could be used together to tease apart the influence of structural markedness versus form-meaning resemblances in driving iconic effects. We tend to assume that iconic effects are 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 should be theoretically justified and matched to the phenomenon the measures will be used to study. For example, if the measures will be used to study language processing in adults, then adult native speakers should be used. 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 tested 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. We found that the former method was much more sensitive to the particular words chosen as translations and foils. This is as anticipated, as the English-speaking participants are expected to have more associations with the English translations than with the unknown Japanese words, leading to more complications in the design where participants guess between the translations, compared to the design where they guess between words.

Not only that, but the spread of results is much wider–with more words that were guessed at high accuracy levels, and less words sitting around chance level–when guessing between words than when guessing between translations. This is also not surprising, as in order to make comparisons between the guessing accuracies fair when guessing between translations, the translations used were restricted to words from the same semantic domain. Not only that, but since some of the words in the data did not have natural antonyms, these could not be antonyms either–otherwise the task would be easier for words with antonyms than words without them. This made the design where participants guessed between translations harder than the design where they guessed between words, as the foil words were designed to be maximally phonologically distinct.

A task where participants guess between words is also simply easier to operationalise, especially when using a design where the foils are chosen randomly from the other trials. When guessing between translations, pains need to be taken to find foils that are from the same semantic domain, are a similar length of characters, and (if your dataset contains any words which do not have natural antonyms) are not synonyms nor antonyms of the correct translation. This makes the guessing between translations design more difficult to use, particularly in small semantic domains (e.g. taste). The guessing between words design, in contrast, can be easily adopted to any semantic domain or indeed any kind of data. Moreover, it is also preferable for being more comparable to the design in the rating task–as in both cases participants are given the meaning first, and only one English translation is involved.

Since the stimuli in this study had artificially flat intonations, which has been shown to decrease performance in guessing studies (Kunihira 1971; Dingemanse et al. 2016), we tried to make the guessing task easier in this study by using foils chosen to be as phonologically distinct as possible from the word being tested, rather than random foils. This was also theoretically justified for a task designed to pick up on iconicity, since if a word sounds suited to its meaning, then a word that sounds very different to it should not sound suited to that meaning. The extra step was worthwhile in this case, as it allowed us to pick up on more iconicity in our small dataset, even with degraded stimuli. In datasets that are relatively higher in iconicity–e.g. gestures or hand signs–or where stimuli are more natural, the extra step may well not be necessary. That is, you may well be able to detect a reasonable amount of iconicity simply by using randomly chosen foils.

We also compared our iconicity ratings from non-speakers with iconicity ratings from native speakers in Thompson, Akita, and Do (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 these words were more likely to be iconic in the first place, and so raters were more likely to have stronger intuitions about these words. 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 use whichever design is most theoretically justified (see discussion in Section 4.1).

## 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 (<https://pypi.org/project/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.

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. <https://doi.org/10.1371/journal.pone.0198430>.

Aryani, Arash, and Arthur M. Jacobs. 2018. “Affective Congruence Between Sound and Meaning of Words Facilitates Semantic Decision.” *Behavioral Sciences* 8 (6): 56. <https://doi.org/10.3390/bs8060056>.

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.

Diffloth, Gérard. 1994. “I: Big, a: Small.” *Sound Symbolism*, 107–13.

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–e133.

Dingemanse, Mark, and Bill Thompson. 2020. “Playful Iconicity: Structural Markedness Underlies the Relation Between Funniness and Iconicity.” *Language and Cognition* 12 (1): 203–24. <https://doi.org/10.1017/langcog.2019.49>.

Hall, Kathleen Currie, J. Scott Mackie, and Roger Yu-Hsiang Lo. 2019. “Phonological CorpusTools: Software for Doing Phonological Analysis on Transcribed Corpora.” *International Journal of Corpus Linguistics* 24 (4): 522–35.

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. <https://doi.org/10.1017/langcog.2016.28>.

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). <https://doi.org/10.1515/lingty-2020-2034>.

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. <https://doi.org/10.1016/S0022-5371(71)80042-7>.

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.

Mielke, Jeff. 2012. “A Phonetically Based Metric of Sound Similarity.” *Lingua* 122 (2): 145–63.

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. <https://benjamins.com/catalog/gest.16.1.04occ>.

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. <https://doi.org/10.3389/fpsyg.2018.00725>.

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. <https://doi.org/10.1038/s41598-018-20961-6>.

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.

Shinohara, Kazuko, and Shigeto Kawahara. 2010. “A Cross-Linguistic Study of Sound Symbolism: The Images of Size.” In, 36:396–410.

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). <https://doi.org/10.1515/lingvan-2019-0088>.

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): 72.

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. <https://doi.org/https://doi.org/10.1111/cogs.12474>.

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. <https://doi.org/10.3389/fpsyg.2016.02073>.

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. <https://doi.org/10.3389/fpsyg.2016.01200>.

Ultan, Russell. 1978. “Size-Sound Symbolism.” In *Universals of Human Language*, edited by Joseph H. Greenberg, 2:525–68. Stanford: Stanford University Press.

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.

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown. 2010. “Sound Symbolism in Basic Vocabulary.” *Entropy* 12 (4): 844–58. <https://doi.org/10.3390/e12040844>.

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. n.d. “Trilled/R/Is Associated with Textural Roughness in the Vocabularies of Spoken Languages.”

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. <https://doi.org/10.3758/s13414-017-1361-2>.

1. 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. [↑](#footnote-ref-1)
2. We excluded translations with fewer than 10 ratings from this analysis as the t-test is less robust against small sample sizes than the fisher test. [↑](#footnote-ref-2)