Do all aspects of learning benefit from iconicity? Evidence from motion capture

Citation for published version:
Sato, A, Schouwstra, M, Flaherty, M & Kirby, S 2019, 'Do all aspects of learning benefit from iconicity? Evidence from motion capture', *Language and cognition*, vol. 12, no. 1, pp. 36-55.
https://doi.org/10.1017/langcog.2019.37

Digital Object Identifier (DOI):
10.1017/langcog.2019.37

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Language and cognition

Publisher Rights Statement:
This article has been published in a revised form in Language and Cognition, doi:10.1017/langcog.2019.37. This version is free to view and download for private research and study only. Not for re-distribution, re-sale or use in derivative works. © Asha Sato, Marieke Schouwstra, Molly Flaherty and Simon Kirby.

Do all aspects of learning benefit from iconicity? Evidence from motion capture.

Asha Sato, Marieke Schouwstra, Molly Flaherty & Simon Kirby

Centre for Language Evolution, University of Edinburgh.

Abstract:

Recent work suggests that not all aspects of learning benefit from an iconicity advantage (Ortega, 2017). We present the results of an artificial sign language learning experiment testing the hypothesis that iconicity may help learners to learn mappings between forms and meanings, whilst having a negative impact on learning specific features of the form. We used a 3D camera (Microsoft Kinect) to capture participants' gestures and quantify the accuracy with which they reproduce the target gestures in two conditions. In the iconic condition, participants were shown an artificial sign language consisting of congruent gesture-meaning pairs. In the arbitrary condition, the language consisted of non-congruent gesture-meaning pairs. We quantified the accuracy of participants' gestures using dynamic time warping (Celebi et. al., 2013). Our results show that participants in the iconic condition learn mappings more successfully than participants in the arbitrary condition, but there is no difference in the accuracy with which participants reproduce the forms. While our work confirms that iconicity helps to establish form-meaning mappings, our study did not give conclusive evidence about the effect of iconicity on production; we suggest that iconicity may only have an impact on learning forms when these are complex.

Introduction

Iconicity is increasingly considered to be a fundamental design feature of language alongside arbitrariness. Under this view, it is suggested that arbitrariness and iconicity play complementary roles in language, with arbitrariness serving to maintain discriminability between items, whilst iconicity helps to bootstrap learning (Dingemanse et. al. 2015). In this paper we investigate whether some aspects of learning linguistic items are helped by iconicity more than others.

There is a broad literature base suggesting that iconicity supports learning with evidence from various domains. Some of the most robust evidence for a positive effect of iconicity on lexical learning comes from experiments in the signed modality in which adult hearing non-signers learn signs from existing sign languages. Lieberth & Gamble (1991) tested learners on their recall of iconic and arbitrary signs from American Sign Language. They report that participants were able to recall the English translation for iconic signs for longer (after 1 week) than for arbitrary signs. Beykirch, Holcolm & Harrington (1990) also find a positive effect of iconicity in a similar design, also using ASL. Maynard, Slavoff & Bovillian (1994) specifically exclude iconic signs from their study, but report that providing a sign's etymology led to better recall than motor rehearsal, which suggests that the perceived motivatedness of a mapping may play a role in learning even arbitrary signs. Baus, Carreiras & Emmorey (2013) report that iconicity enhances new learners' (but not proficient signers') performance on a translation task between ASL and English. Morett (2015) shows that iconic signs are learned more effectively than arbitrary or metaphoric signs from ASL[1]. These studies taken together suggest that iconicity is helpful to adult learners in the signed modality, using real signed languages.

Imai & Kita (2014) have proposed more generally that non-arbitrary relationships between forms and meanings - in particular sound symbolic mappings- form an essential design feature of language with an important role in acquisition. There is evidence to support the idea that sound-symbolism confers a learning advantage from studies with real languages (see e.g. Imai, Kita Nagumo & Okada, 2008; Nygaard, Cook & Namy, 2009; Lockwood & Dingemanse, 2016), and artificial languages (Perlman & Lupyan, 2018), as well as a large body of work suggesting a bias towards sound-symbolic mappings in tasks that don't involve

---

[1] Metaphoric signs are described by the author as signs depicting concrete attributes associated with an (abstract) referent metaphorically, for example GOAL, which involves moving one hand toward the other with the index finger extended. They are therefore related to iconic signs in that they are motivated signs.

learning (see e.g. Nielsen & Rendall, 2012, Hirata et. al., 2011). A more comprehensive review of research on sound symbolism can be found in Lockwood & Dingemanse (2015).

It is worth noting that studies with children in both spoken and signed modalities generally present more of a mixed picture than work with adults. Although there are indications that iconicity plays a role in early language learning (Thompson, Vinson, Woll & Vigliocco, 2012; Perry, Perlman & Lupyan, 2015), some authors have found a null or negative effect of iconicity (e.g. Fort at. al. 2013), suggesting there may be additional factors to take into consideration when considering child learners. In the signed modality, one such consideration is motor development, which may pose articulatory constraints such that children's errors result in lower iconicity (Meier, Mauk, Cheek & Moreland, 2008). Young children may also have difficulty interpreting or recognising iconic links (Tolar, Lederberg, Gokhale & Tomasello, 2008), some of which may be culturally specific and therefore dependent on experience (Occhino, Anible, Wilkinson & Morford, 2017). Children may also have biases for different types of iconicity that differ from those of adults (Ortega, Sumer & Özyürek, 2017).

The best evidence for a positive effect of iconicity in learning then comes from work with adults, but despite the apparently robust finding that iconicity supports the learning of form-meaning pairs where adult learners are concerned, some recent studies suggest that this contribution to learning may not be a straightforward advantage: In an artificial language learning experiment using a whistled language, Verhoef et. al. (2016) found that whistles were reproduced with higher error in a condition where iconicity was possible compared to a condition where iconicity was disrupted by scrambling the mapped correspondence between signals and meanings. Similarly, in a longitudinal study of phonological development in adult British Sign Language learners, Ortega & Morgan (2015) found that iconic signs are articulated less accurately than arbitrary signs of equal complexity after 11 weeks of classes. These two results are potentially at odds with the idea that iconicity supports learning, but we are in agreement with Ortega (2017) that this divergence in results reflects a methodological difference in precisely what is reflected in the task used to measure of learning: Reviewing the effect of iconicity in the sign language learning literature, Ortega (2017) convincingly shows that studies that find a positive effect of iconicity use tasks that relate to the meaning of the sign (like the translation task in Baus et. al. (2012)), whilst only those studies that use measures relating to phonological[2]

---

[2] I follow Ortega's terminology here.

aspects of signs find null (Morett, 2015) or negative effects (Ortega & Morgan, 2015)[3]. He concludes that iconicity assists with the semantic aspects of L2 lexical sign acquisition, but not the phonological aspects. This observation also appears to be true for other modalities, including in artificial language learning, and studies on sound symbolism using real words from spoken languages. In these kinds of studies, tasks often reflect the mapping between the target form and meaning (c.f. Ortega's semantic aspects), rather than specific properties of the form (c.f. phonological aspects). For example, learners might be required to state whether a form-meaning pair was in the training set (e.g. Lockwood, Dingemanse & Hagoot, 2016), or pick an item from an array in a forced choice paradigm (e.g. Nygaard, Cook & Namy, 2009). These tasks establish whether a mapping between form and meaning has been learned, but do not provide insight into learners' productions of the forms themselves.

To summarise, the prevalence of use of measures based on the mapping between form and meaning can potentially obscure effects relating to the learning of the form itself. The aforementioned Verhoef et. al. (2016) cite an example from their experiment in the discussion which we think is reveals the relevance of differentiating between the mapping and the form, showing two whistles which map iconically onto the referent in the same way (and would presumably have led to success in a task based on the mapping), but are formally dissimilar to one another, as the whistle pitch contours are mirror images of one another. This example from Verhoef et. al. is suggestive of a strategy for encoding iconic signals which makes use of the potential to reproduce a form on the basis of its presumed resemblance to its referent, as opposed to a strategy in which the specific features of the form are internalised. In other words, if a learner does not perceive the form to be iconic they do not know which features of the form are relevant, and therefore must give all aspects of the form equal weight. On the other hand, when learners do perceive a form to be iconic, their strategy for reproduction may involve reconstructing a quite different form that maintains only the iconic relation to the referent. Indeed, the use of formally different strategies to encode iconic meanings has been reported in an emergent sign language (Sandler et. al., 2011). The existence of various strategies for iconically encoding a referent could create precisely the kind of variability reported by Verhoef et. al. and Ortega & Morgan. This variability is to be expected *especially* in cases where learners already have a form in their communicative repertoire which corresponds to the referent: Ortega & Özyürek (2013) show that learners' articulation of iconic BSL signs overlap significantly with their own gestural repertoire. Ortega,

---

[3] Although Morett (2015) uses a phonological measure, she finds no difference in recall or accuracy of iconic, arbitrary and metaphoric signs when testing immediately after learning (delay of 5 mins), but reports a positive effect of iconicity when re-testing after 1 and 4 weeks.

Schiefner & Ozyurek (2019) additionally show show that learners' accuracy guessing the meaning of signs as well as their perception of iconicity in Sign Language of the Netherlands (NGT) is predictable from pre-existing gestural repertoires. Iconic links can be established on the basis of existing communicative forms, however such an explanation is unlikely in the case of artificial languages like the whistled language used by Verhoef et. al.

On this basis, we aim to test the hypothesis that iconicity may contribute differently to the learning of the different component parts of linguistic items. Specifically, we predict that iconicity may help learners to learn a mapping, whilst having a negative effect on learning of the exact properties of the form. We test this hypothesis with an artificial sign language learning experiment. By using the term artificial sign language we do not intend to suggest that the necessarily simplified system of gestures we test in our experiment are comparable to the full complexity of natural signed languages. Instead, we use the term to indicate that our methodology is an extension of artificial language learning paradigms (see Motamedi, Schouwstra, Smith, Culbertson & Kirby, 2019). That is, our stimuli represent a necessarily simplified system, aiming to isolate the particular properties we are interested in, and we hope that the result is language-like in relevant and informative ways. We use depth and motion tracking technology (Microsoft Kinect) to quantify the accuracy with which learners reproduce target gesture forms. This kind of technology has been identified as a key new tool in the analysis of sign and gesture, comparable to the spectrograph for spoken language (Goldin-Meadow and Brentari, 2017). The Kinect in particular has been used by Namboodiripad, Lenzen, Lepic & Verhoef (2016) to show a reduction in the size of the articulatory space as gestures become conventionalised.

Methods

Participants

We recruited 38 participants to take part in an artificial sign language learning experiment. Participants were right handed, and reported having no previous knowledge of any signed language.

Stimuli

We created an artificial sign language consisting of 32 iconic, pantomimic gesture-meaning pairs (see figure 1). The set was designed so that the iconicity of each gesture was predominantly contained in the movement (as opposed to e.g. the handshape), as this is the parameter most suitable for quantification using the Kinect.

The gestures included items such as “to brush hair”, paired with three vertical movements on one side of the head, and “to push”, paired with a single outward sagittal movement with both hands. Video files containing the 32 gestures and example training languages can be found in the supplementary materials.

Procedure

Each participant was trained on 16 gesture-meaning pairs in a between subjects design. In the iconic condition, participants saw 16 iconic gesture-meaning pairs, selected at random from the set of 32. In the arbitrary condition, participants also saw 16 gestures, but these were paired with the 16 meaning labels from the non-selected set (see figure 2). This created random pairings between gesture forms and meanings, and avoided the possibility of a gesture appearing in the same set as the label it would have been paired with in the iconic condition. The gestures were presented as 3D stick figure animations, recorded using a Microsoft Kinect depth and motion sensor and smoothed using median filtering to remove jitter (Microsoft, 2014).

The experiment started with a familiarisation phase in which participants acquainted themselves with the sensor. Participants completed 4 rounds of training trials in which they guessed the meaning of each gesture (with feedback) before copying it, followed by a test stage in which they saw a label and produced a gesture from memory, and finally an iconicity rating stage. Participants’ body movements were recorded using the Kinect throughout the training and test trials.

In the familiarisation phase, participants completed a set list of instructions (wave, touch your nose, put your hands on your head, put your hands on your hip) to familiarise them with the type of information the device records. They were also told explicitly that the figure was mirrored, rather than rotated, so that the right side of the figure on the screen corresponded to the right side of the participant’s body.

In each training trial, the participants first saw an animation of the 3D figure performing the target gesture. They were then asked to guess the meaning of the gesture, choosing from an array containing the correct mapping and two distractor items. Distractor items were randomly selected from the set of meanings within the participant’s training set. They were given feedback on their guess before seeing the animation a second time, paired with the target label, and were then asked to imitate the gesture. Participants saw a 3 second count-down followed by a “RECORDING” screen, and were prompted to indicate verbally to the experimenter when they were finished producing the gesture by saying “next”, at which point the experimenter started the next trial. In each training round, participants saw all 16 items in a randomised order.

In the test stage, participants saw all 16 target labels on the screen, and were able to select the order in which they wanted to record their responses, though they were ultimately required to provide a gesture for each label. The procedure for recording was the same as in the training trials, with a 3 second count-down followed by a “RECORDING” screen and a prompt to indicate to the experimenter when they were ready to move on with the word “next”.

After the test stage, we obtained iconicity ratings by asking participants to rate each gesture according to how much they thought it resembled its meaning. They were provided with an example as follows: *“If you saw a gesture that meant ‘digging’ and the gesture resembled a digging action, you would give it a high rating. If on the other hand the gesture for ‘digging’ did not resemble a digging action, you would give it a low rating.”* They were asked to give two ratings for each gesture-meaning pair that they saw. The first iconicity rating asked participants to rate the iconicity of the gesture-meaning pair based on their initial impression (on first exposure). The second iconicity rating asked participants for their post-hoc impression, having produced the gestures several times during the experiment, and the instructions noted that the second rating need not be different to the first. Participants gave their ratings on an 7 point Likert scale using two sliders. They were prompted with the question *“How much did the gesture resemble its meaning?”*, and were instructed to use the first slider to give their first impression and the second slider to give their post-hoc impression. The ends of the scales were labelled *“completely”* and *“not at all”*, and participants were free to choose a point anywhere along the scales.

Results

Iconicity ratings

The iconicity ratings given by participants show that those in the iconic condition indeed considered that the items they saw were iconic, and those in the arbitrary condition perceived the items they saw as arbitrary, both in their first impression (iconicity rating 1) and their post-hoc rating (iconicity rating 2) (figure 3). The decision to use two scales was based on reports from participants in a pilot experiment that their first impression was different to the rating they would give post-hoc. We therefore expected participants might use the two scales differently, but did not have specific hypotheses about how they would do so. The data showed an overall increase in the perception of iconicity in the second, post-hoc rating.

We ran a mixed effects linear model predicting iconicity ratings with fixed effects of condition (arbitrary or iconic) and rating (iconicity rating 1 or iconicity rating 2), with random intercepts per participant. Likelihood ratio testing confirms our model performs significantly better than the null model with random effects only ($X^2(2) = 104.32$, $p <.001$). Iconicity ratings were higher by 2.4 (s.e. 0.19) in the iconic condition relative to the model intercept of 2.9 (s.e. 0.15), with an additional increase of 0.6 (s.e. 0.08) for iconicity rating 2. Comparisons to reduced models with each fixed effect removed confirm a main effect of condition ($X^2(1) = 59.6$, $p < .0001$) and rating ($X^2(1) = 44.6$, $p < .0001$).

Although the iconicity ratings reflect the intended classification, there was some variation, with some items being rated as completely arbitrary by participants in the iconic condition, and vice versa. The following analyses were therefore computed in two ways, first by condition (iconic/arbitrary), and then using participants' own iconicity ratings. For the analyses using participants' own iconicity ratings, we report the results using the first iconicity rating (reflecting participants' first impression). It is possible that the difference we observed between the rating of the first impression and the pos-hoc impression could be explained by the salience of asking for two ratings: participants may have been providing the answers they believed the researchers were looking for. However, results using the second iconicity rating do not differ from those presented below in terms of direction or significance.

Learning of the mapping

In order to assess participants' learning of the mapping between form and meaning, we looked at how often participants responded correctly to the guessing task that formed part of the training trials. As expected, iconicity had a positive effect on learning the mapping (figure 4). In the first training round, participants in the iconic condition performed better than participants in the arbitrary condition on the guessing task. Note that in the first training round, participants have not yet been exposed to the mapping, so these responses represent a naïve guess. Nonetheless, participants in the iconic condition guessed correctly 94% of the time, presumably reflecting the transparency of the iconic gesture-meaning pairs. It is worth noting that participants in the arbitrary condition also performed above chance in the first round, guessing correctly 56% of the time (one sample t-test, $t = 9.1463$, $df = 18$, $p < .001$). Because the correspondences between gestures and labels were assigned randomly for each participant in the arbitrary condition, it is unlikely that this success rate is due to accidental iconicity in the gesture-meaning pairs.[4] Instead, we suggest the high success rate in the first training round is likely to be the result of participants using a process of elimination to discount labels for items they have already learned as the round progresses. This is similar to what Smith, Smith & Blythe (2010) call cross-situational learning.

We ran a logistic regression predicting success on the guessing task with fixed effects of condition (iconic/arbitrary) and round (1 to 4), as well as their interaction, with random intercepts per participant (adding random slopes led to convergence issues). Likelihood ratio testing confirms the fitted model to perform significantly better than the null model with random effects only ($X^2(3)= 256.21$, $p < .0001$). According to the model, the likelihood of giving a correct answer is higher in the iconic condition (odds ratio 10.66), and higher with subsequent rounds (odds ratio 2.49). The interaction between condition and round was not significant as tested by comparison of the full model to a model with the interaction removed ($X^2(1) = 1.23$, $p = 0.27$). Comparison to reduced models with each fixed effect removed confirms a main effect of both condition ($X^2(1)= 63.88$, $p < .0001$) and round ($X^2(1)= 188.10$, $p < .0001$).

These findings are corroborated in a model using participants' own iconicity ratings; we ran a mixed effects model predicting success on the guessing task with a fixed effect of iconicity rating, and random intercepts per participant. Likelihood ratio testing confirms the fitted model performs better than the null model with random effects only ($X^2(1)= 61.37$, $p < .0001$). According to the model, the likelihood of giving a correct answer increases per unit increase in iconicity rating (odds ratio 0.36).

Learning of the form

We measured the accuracy with which learners copied the gesture forms by comparing the trajectory of the wrists during gesture production to the trajectory of the wrists in the model gesture. In order to do this, we first transformed each participant's body-tracking data, scaling the size of their tracked skeleton to that of the model gesture, using a scaling factor obtained from the length of the upper arm (average euclidean distance

---

[4] Although we can not assume that participants do not perceive iconicity in the random pairings (Emmorey, 2018, see also Occino, Anible, Wilkinson & Morford, 2017), and our participants indeed sometimes perceive the scrambled pairs as highly iconic, it is still unlikely that the target arbitrary pairings would systematically be perceived iconically more often than the pairings with the distractor items, as they are drawn randomly from the same pool.

between the tracked shoulder and elbow joints), and applied the same median filter smoothing as was applied to the target gestures. We compared the trajectory of the wrists using dynamic time warping (DTW) (Celebi , Aydin, Temiz & Arici, 2013; see also Verhoef et. al., 2015), a technique for determining the similarity between sequences of time series data which can be used as a distance measure. In this case the reference sequence contains the tracked XYZ position of the model's left and right wrists in each frame, and the test sequence contains the XYZ position of the participant's left and right wrists (after scaling and filtering) for the frames in the participant's response. The XYZ coordinates are measured in cm from the position of the tracked joint between the shoulders. We report the computed DTW distance, normalised for path length (Giorgino, 2009). In order to guide the reader in interpreting the results that follow from this perhaps unintuitive measure, figure 5 shows the DTW distance computed between pairs of model gestures that are impressionistically similar (*to clean* and t*o knock on a door,* yielding a normalised DTW distance of 36.22) and impressionistically dissimilar (*to clean* and *to swim* yielding a normalised DTW distance of 219.57). In order to assess the suitability of this measure for our analysis we first checked that the distance between participants' produced gestures and the target item was reliably smaller than the distance to non-target items. For this we compared each gesture that each participant produced (after scaling and filtering) to the target model gesture, and to each of the 31 non-target model gestures (figure 6). A one sample t-test confirms that the distance to target items was reliably smaller ($t = 52.24$, $df = 3411.90$, $p < .001$). Despite a clear positive effect of iconicity on learning the mapping, iconicity had (counter to our prediction) no effect on the learning of the form: there was no difference in copying accuracy between the iconic and arbitrary conditions in either the training rounds or the test stage (figure 7).

In the training rounds, the mean normalised DTW distance to the target item in the iconic condition was 125.11, and in the arbitrary condition was 132.14. We ran a mixed effects linear model predicting the normalised DTW distance to target gesture with a fixed effect of condition and random intercepts per participant. Model comparison by likelihood ratio testing shows that the fitted model is not significantly better than the null model with random effects only ($X^2(1)= 2.62$, $p = .10$), revealing no relation between iconicity and copying accuracy during training.

Using participants' own ratings, we ran a mixed effects linear model predicting normalised DTW distance to target gesture with a fixed effect of iconicity rating, and random intercepts per participant. This model is significantly better than the null model with random effects only ($X^2(1)= 7.23$, $p =.007$) and shows and effect in the opposite direction to our prediction, revealing that for items with high perceived iconicity, participants had higher copying accuracy during training. However, the size of the effect is small: an estimated decrease in nDTW of 1.01 (standard error 0.36) per unit increase in iconicity from the model intercept of 133.30.

In the test round, the mean normalised DTW distance to the target item for participants in the iconic condition was 129.50, and in the arbitrary condition 143.72. We ran a mixed effects linear model predicting the normalised DTW distance to target gesture with a fixed effect of condition and random intercepts per participant. Model comparison by likelihood ratio testing shows that the fitted model is not significantly better than the null model, again revealing no relation between iconicity and accuracy in reproducing the form at test ($X^2(1)= 5.58$, p = .02).

Using participants' own iconicity ratings, a mixed effects linear model predicting normalised DTW distance to target gesture with a fixed effect of iconicity rating and random intercepts per participant performs significantly better than the null model with random effects only ($X^2(1)= 10.69$, p = .001). Again, the effect is in the opposite direction to our prediction, suggesting that participants were more accurate producing items with high perceived iconicity at test. However, as above, the size of the effect is small, with an estimated decrease in nDTW of 2.53 (standard error 0.77) per unit increase in iconicity rating, from the model intercept of 147.62.

Consistency of form

In order to assess how consistently participants produced each gesture, we used the same distance measure described above (normalised DTW) to carry out pairwise comparisons of the 4 tokens of each gesture-meaning pair produced by each participant during the training rounds. We report the mean distance between tokens (figure 8). The mean distance between tokens for participants in the iconic condition was 55.28, and for participants in the arbitrary condition was 58.65.

We ran a mixed effects linear model predicting the mean distance between tokens with a fixed effect of condition and random intercepts per participant. Comparison to the null model by likelihood ratio testing suggests that the fitted model does not perform better than the null model with random effects only ($X^2(1)=$ 0.45, p = 0.50).

Using participants' own iconicity ratings, we ran a mixed effects linear model predicting the mean distance between tokens with a fixed effect of iconicity rating, and random intercepts per participant. This model performs better than the null model with random effects only ($X^2(1)= 9.37$, $p = 0.002$), showing an effect in the opposite direction to our prediction, though again the effect size is small. The model estimates a decrease in distance between tokens of 0.81 (standard error 0.27) per unit increase in iconicity rating, from the model intercept of 60.29.

Discussion

In line with our prediction, and with previous research on the effect of iconicity with adult learners, we found that iconicity helps learners establish a mapping between form and meaning: participants' accuracy at the guessing task was higher in the iconic condition, even before being informed of the intended mapping in round 1, and performance on the guessing task quickly reached ceiling in the iconic condition. Perlman and Lupyan (2018) report a similar pattern of results in their learning task with novel vocalizations, with an early advantage for iconic vocalizations that quickly reaches ceiling. Our finding adds to the existing body of work in different modalities, converging to suggest that iconicity helps learners establish mappings between forms and meanings.

However, counter to our prediction, we did not find any effect of iconicity on the learning of the form. We had hypothesised that when learners perceive a form to be iconic, their strategy for reproducing it may involve reconstructing a form that maintains the iconic relation to the referent but ends up being potentially different, as in the example given from Verhoef el. al. involving a mirrored signal. Our findings do not support this idea, as participants reproduced gestures with similar accuracy in both conditions. Our findings therefore diverge from the negative effect of iconicity observed by Ortega & Morgan (2015) and Verhoef et. al. (2015) that motivated the present study. In what follows we will consider the potential reasons for the difference, and formulate new ideas about the relation between iconicity and complexity.

Returning to the results of the two previous studies, we can observe that in Ortega & Morgan (2015), higher articulatory error for iconic items is significant for items above a complexity level 4 in their 6 level classification (extended from Battison's 1978 classification), but for complexity levels 2-4 there is no significant difference between the iconic and arbitrary items, and for the simplest signs of level 1, iconic items are in fact articulated more accurately. It is worth noting that their measure of articulatory accuracy is based on the parameters of handshape, location, movement and orientation. Of these, only location and movement are reflected by our Kinect-based measure, and we did not systematically vary the complexity of the gestural forms that made up our artificial language. In Verhoef et. al. (2015), higher error in the scrambled condition (in which iconicity is possible) is only seen in the first few generations. By the end of the transmission chain (of 9 generations), the intact and scrambled conditions have converged on similar levels of error. It is not clear how the individual whistles change in their complexity as transmission proceeds, as the study focuses on the emergence of combinatoriality in the sets of whistles. However, we might guess that whistles were more complex in earlier generations, as iterated learning generally leads to systems that are simpler and more expressive (Kirby, Cornish & Smith, 2008). Therefore, we can tentatively relate the partial negative effect of iconicity on the learning of forms in these studies to differences in the complexity of the items involved. Importantly, a negative effect of iconicity was not explicitly part of the prediction in either study, and indeed that was our motivation for running the present study.

This leads us to consider the following option: might it be the case not so much that our results are inconsistent with those in the studies which motivated this one, but that our prediction was formulated in a way that missed something important about the patterns of results in those studies, specifically the role played by the complexity of the items? Although we can not draw any conclusions post-hoc, it may be that iconicity can both help and hinder the learning of forms, depending on the complexity of the item. It may be, for instance, that for simple items, iconicity helps learners to perceive basic components of the gesture, such as the general direction of a movement. Rosen (2004) reports that the parameter most difficult for adult L2 learners of American Sign Language to perceive is movement. Consistent with this, we observe cases in the arbitrary condition of our experiment where participants apparently have difficulty parsing the direction of the movement in very simple target gestures. In one example a participant produces an upward vertical movement in response to an item where the target movement is forward and horizontal. In the iconic condition, the same target gesture is paired with the label *to push*, and it is easy to see how for this item iconicity could help the learner parse the direction of the movement in a way that would not happen for an arbitrary pairing. Indeed, we observed no visibly comparable errors in the iconic condition. Given that the direction of the movement is perhaps the most relevant feature for this gesture, it seems like a case where iconicity could help learning of the form, whilst for more complicated items involving more complicated movement, or additional parameters such as handshape (not manipulated in our stimuli), iconicity could still lead to errors either based on the learner's own experience (such as their pre-existing gestural repertoire), as described by Ortega, Schiefner & Ozyurek (2019) or based on lower attention to detail when faced with iconic mappings.

If there is an interaction such that a negative effect of iconicity on forms emerges only for complex items, it is probable that the pantomimic gestures in our study fell within a range of complexity that would not have elicited this effect. It is difficult to make any direct comparison regarding the complexity of our stimuli and the set of whistles used by Verhoef et. al. due to the difference in modality (and the affordances of different modalities may well be relevant in determining whether iconicity has an effect on learning forms) but impressionistically our target gestures appear on average to have fewer segments (change in direction being perhaps the most comparable way of segmenting continuous movement) than the starting set of whistles used by Verhoef et. al.. In comparison to the BSL signs used by Ortega & Morgan, our target gestures were necessarily simpler, handshape and orientation not being manipulated in our stimuli. In addition to this overall relative simplicity, most of our target gestures were either one-handed, or symmetrical with both hands performing the same movement. Only 4 of our 32 target gestures were two-handed and asymmetrical. The majority of our target gestures would therefore correspond to the lower levels of complexity in the classification used by Ortega & Morgan. Future research could address this idea by expanding the range of complexity of the gestures, and could give more careful consideration to the role of modality.

It is also possible that iconicity simply does not have a negative effect on the learning of forms. However, if this is the case, it should be noted that we still also lack strong evidence for a *positive* effect of iconicity on the learning of forms. The pattern of results obtained is not what we would expect if the iconicity learning advantage were a global effect, applying to the process of learning linguistic items as a whole. . We suggest that more research is needed in this area to further elucidate the contribution of iconicity to learning, and in particular how this contribution may differ for the component parts of linguistic items. Although it is clear that iconicity helps to establish mappings between forms and meanings, it remains unclear whether iconicity has an effect on the precision with which learners reproduce forms, and if so, what factors modulate it. There is scope for future research to further address this question using the Kinect with a stimuli set that is designed to take the complexity of individual items into account.

Supplementary materials:

https://github.com/AshaSato/Do_All_Aspects_Of_Learning_Benefit_From_Iconicity

References

Baus, C., Carreiras, M., & Emmorey, K. (2013). When does iconicity in sign language matter?. *Language and cognitive processes*, *28*(3), 261-271.

Celebi, S., Aydin, A. S., Temiz, T. T., & Arici, T. (2013). Gesture Recognition using Skeleton Data with Weighted Dynamic Time Warping. In *VISAPP* (1) (pp. 620-625).

Dingemanse, M., Blasi, D.E., Lupyan, G., Christiansen, M.H. and Monaghan, P., (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19(10), pp.603-615.

Giorgino T (2009). "Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package." *Journal of Statistical Software*, *31*(7), pp. 1-24. <URL: http://www.jstatsoft.org/v31/i07/>.

Goldin-Meadow, S., & Brentari, D. (2017). Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and Brain Sciences, 40*, E46. doi:10.1017/S0140525X15001247

Emmorey, K (2018). Iconicity in Sign Language. Plenary at Evolang conference, Torun.

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

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10681–10686.

Köhler, W. (1947). Gestalt Psychology; an Introduction to New Concepts in Modern Psychology, Rev. Edn. Oxford: Liveright.

Laing, C. (2018). A role for onomatopoeia in early language: evidence from phonological development. *Language and Cognition*, 1-15.

Lockwood, G. and Dingemanse, M., (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in psychology*, 6, p.1246.

Lockwood, G., Dingemanse, M., & Hagoort, P. (2016). Sound-symbolism boosts novel word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(8), 1274-1281.

Meier, R. P., Mauk, C. E., Cheek, A., & Moreland, C. J. (2008). The form of children's early signs: Iconic or motoric determinants?. *Language learning and development*, *4*(1), 63-98.

Microsoft (2014). Skeletal Joint Smoothing White Paper. {online} available: http://msdn.microsoft.com/en-us/library/jj131429.aspx. Accessed May 2015.

Morett, L. M. (2015). Lending a hand to signed language acquisition: enactment and iconicity enhance sign recall in hearing adult American Sign Language learners. *Journal of Cognitive Psychology*, 27(3), 251-276.

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964.

Namboodiripad, S., D. Lenzen, R. Lepic, and T. Verhoef. (2016). Measuring conventionalization in the manual modality. *The Journal of Language Evolution* 1(2), 109–118.

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, *112*(1), 181-186.

Occhino, C., B. Anible, E. Wilkinson, and J. Morford. (2017). Iconicity is in the eye of the beholder. *Gesture* 16(1), 99–125.

Ortega, G. and Morgan, G. (2015). Phonological development in hearing learners of a sign language: The influence of phonological parameters, sign complexity, and iconicity. *Language Learning*, 65(3), pp. 660-688.

Ortega, G., & Özyürek, A. (2013). Gesture−sign interface in hearing non−signers'first exposure to sign. In Proceedings of the Tilburg Gesture Research Meeting (pp. 1–5). Tilburg, Netherlands: Tilburg University.

Ortega, G., Schiefner, A., & Ozyurek, A. (2017). Speakers' gestures predict the meaning and perception of iconicity in signs. In G. Gunzelmann, A. Howe, & T. Tenbrink (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (CogSci 2017) (pp. 889-894). Austin, TX: Cognitive Science Society.

Ortega, G., Sümer, B., & Özyürek, A. (2017). Type of iconicity matters in the vocabulary development of signing children. *Developmental psychology*, *53*(1), 89.

Perry LK, Perlman M, Lupyan G (2015) Iconicity in English and Spanish and Its Relation to Lexical Category and Age of Acquisition. PLoS ONE 10(9): e0137147.

Perniss, P., Thompson, R. and Vigliocco, G., (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in psychology*, 1, p.227.

Rosen, R. S. (2004). Beginning L2 production errors in ASL lexical phonology: A cognitive phonology model. *Sign Language and Linguistics*, *7*, 31–61.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross−situational learning: An experimental study of word−learning mechanisms. *Cognitive Science*, *35*(3), 480-498.

Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M., & Pfau, R. (2015). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences*, 112(19), 5968 LP-5973.

Thompson, R.L. Vinson, D. P., Woll, B. & Vigliocco, G. (2012). The Road to Language Learning Is Iconic: Evidence From British Sign Language. Psychological Science.

Verhoef, T., Kirby, S., & Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive science*, 40(8), 1969-1994.

Figure 1: Example gestures and labels.

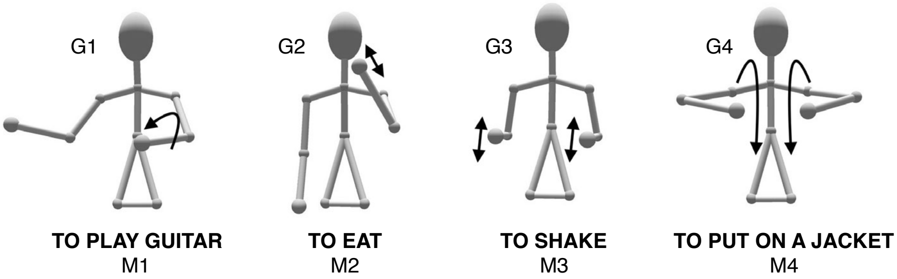


Figure 2: In the arbitrary condition, 16 of the 32 gestures are selected, and are randomly allocated labels from the non-selected set. This ensures that the label G1 *to play guitar* would never appear in the same arbitrary language as the gesture G1 *to play guitar,* avoiding languages that were actively misleading.

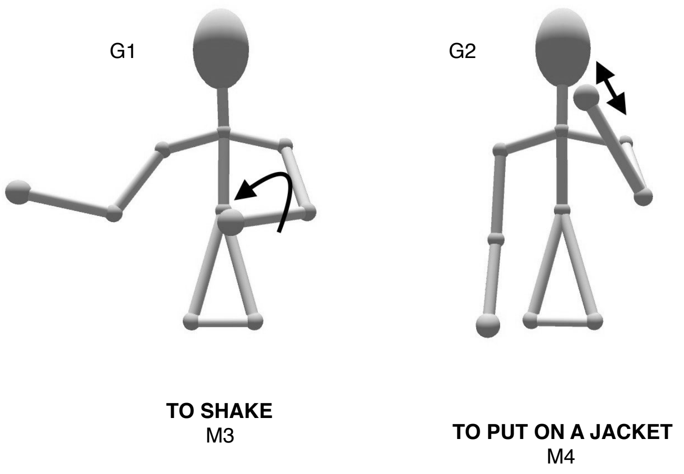


Figure 3: Participant iconicity ratings based on their first impression at first exposure to the mapping (iconicity rating 1) and post hoc impression having completed the experiment (iconicity rating 2).

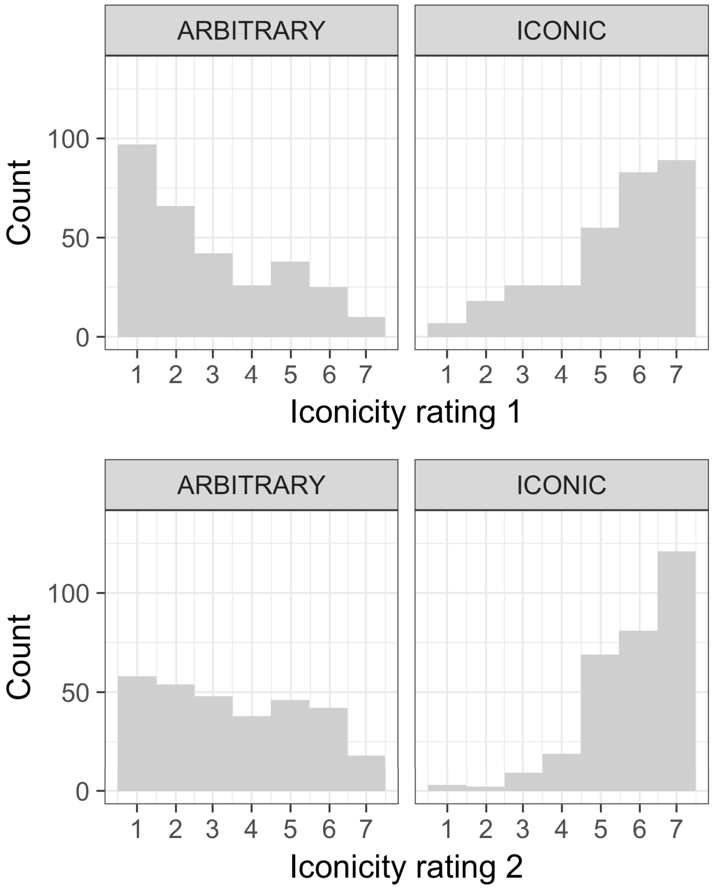


Figure 4: Learning of the mapping: Proportion of correct responses in pre-imitation guessing task in the two conditions. Dashed line represents chance.

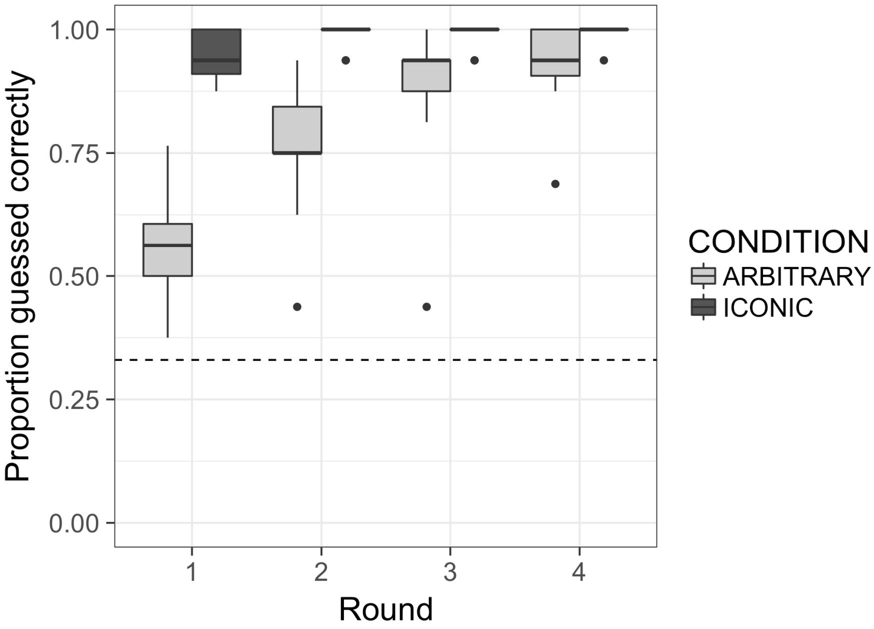
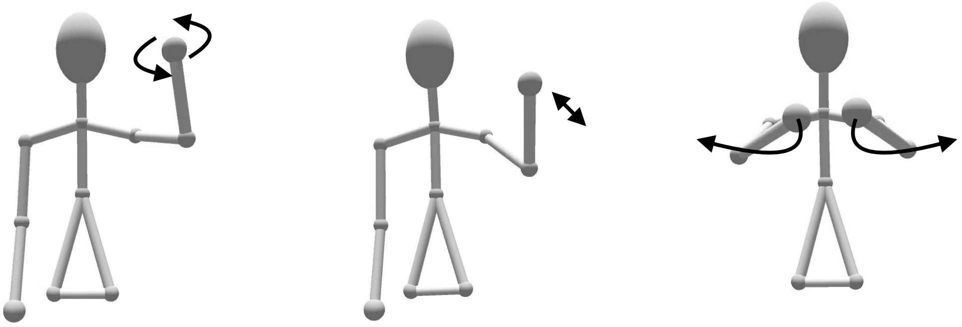


Figure 5: The normalised DTW distance between the impressionistically similar model gestures *to clean* and *to knock on a door* is 36.22. The distance between *to clean* and the impressionistically dissimilar gesture *to swim* is 219.57.



TO CLEAN

TO KNOCK ON A DOOR

TO SWIM

Figure 6: normalised dynamic time warping distance to the model gesture for target and non-target items.

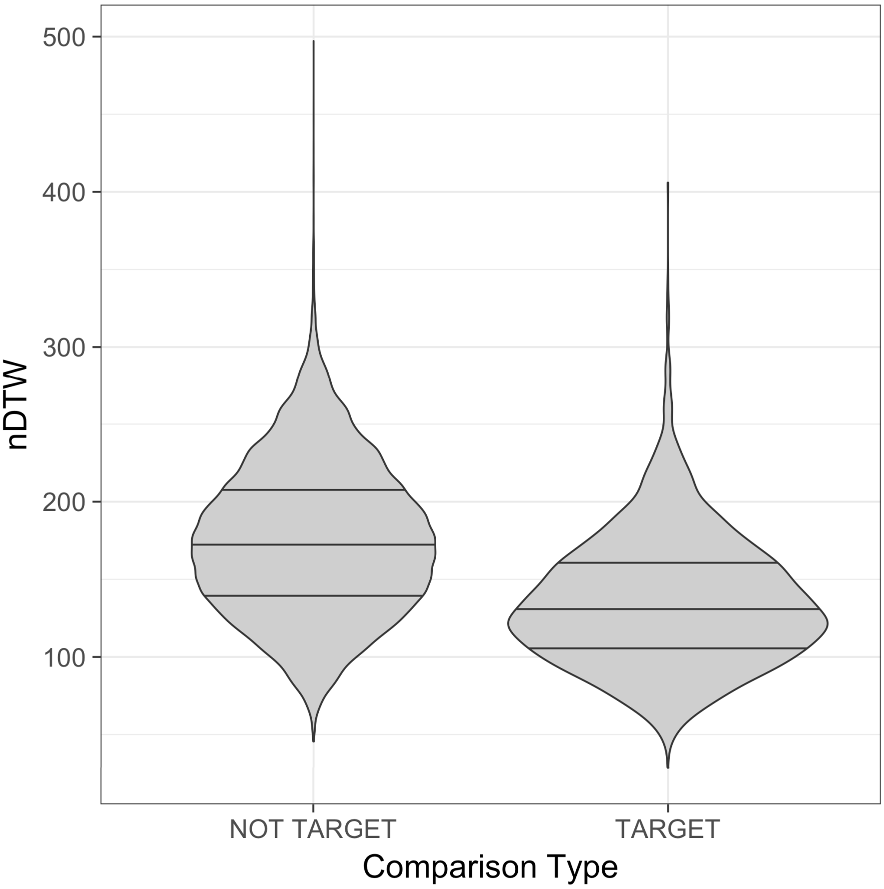


Figure 7: Accuracy copying the form. Mean nDTW to target gesture by condition, and using participants’

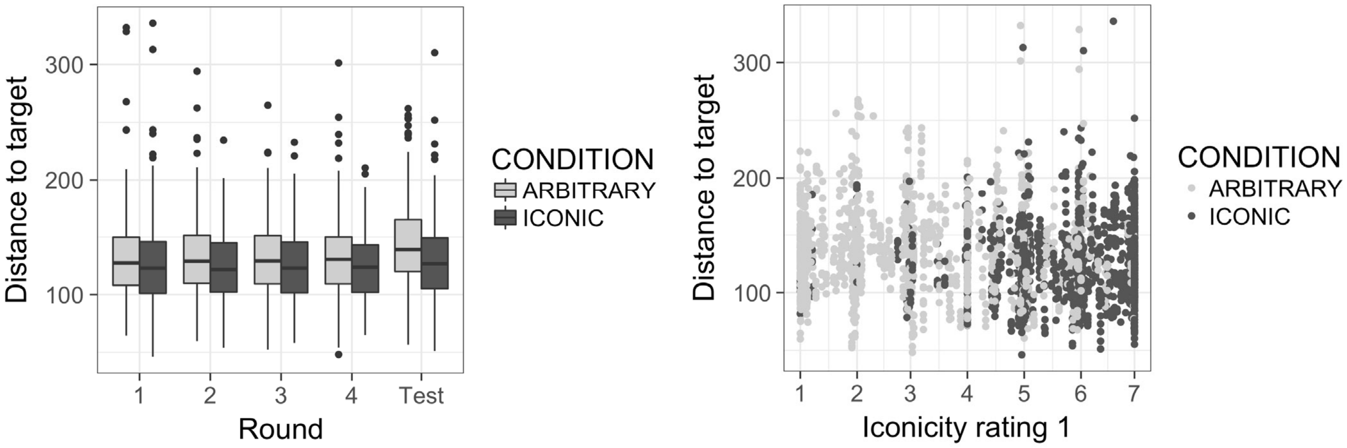


Figure 7: Consistency producing the form. Mean nDTW distance between the 4 tokens produced by each participant during training, by condition and using participants' own iconicity ratings.