

# **Iconicity and Spoken Language**

**John Matthew Jones**

**Thesis submitted in partial fulfilment of the requirements for the degree**

**of**

**Doctor of Philosophy**

**in Psychology and Language Science**

**UCL**

**2016**



## **Declaration**

I, John Matthew Jones, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.



## Abstract

Contrary to longstanding assumptions about the arbitrariness of language, recent work has highlighted how much *iconicity* – i.e. non-arbitrariness – exists in language, in the form of not only onomatopoeia (*bang*, *splash*, *meow*), but also sound-symbolism, signed vocabulary, and (in a paralinguistic channel) mimetic gesture. But is this iconicity ornamental, or does it represent a systematic feature of language important in language acquisition, processing, and evolution?

Scholars have begun to address this question, and this thesis adds to that effort, focusing on spoken language (including gesture). After introducing iconicity and reviewing the literature in the introduction, Chapter 2 reviews sound-shape iconicity (the “kiki-bouba” effect), and presents a norming study that verifies the phonetic parameters of the effect, suggesting that it likely involves multiple mechanisms. Chapter 3 shows that sound-shape iconicity helps participants learn in a model of vocabulary acquisition (cross-situational learning) by disambiguating reference. Variations on this experiment show that the round association may be marginally stronger than the spiky, but only barely, suggesting that representations of lip shape may be partly but not entirely responsible for the effect.

Chapter 4 models language change using the iterated learning paradigm. It shows that iconicity (both sound-shape and motion) emerges from an arbitrary initial language over ten ‘generations’ of speakers. I argue this

shows that psychological biases introduce systematic pressure towards iconicity over language change, and that moreover spoken iconicity can help bootstrap a system of communication.

Chapter 5 shifts to children and gesture, attempting to answer whether children can take meaning from iconic action gestures. Results here were null, but definitive conclusions must await new experiments with higher statistical power.

The conclusion sums up my findings and their significance, and points towards crucial research for the future.

## **Acknowledgments**

Thank you: Gabriella Vigliocco for patient supervision; Joe Devlin and Linda B. Smith for patient secondary supervision; Dave Vinson, Pamela Perniss, Nourane Clostre, Alex Lau-Zhu, Julio Santiago, and Char Wozniak for (in)valuable help and support; Kari Vyas, Leila Jameel, Tom Hardwicke, Joe Walmswell, John Lynch, Will Vaughan, and the youth of Bloomington for friendship and the extracurricular; Lawrence Hill-Cawthorne for setting a good example; Zoë Belk for curing me of dendrophobia; Caroline Lehr for visiting the pirates; and all of my family – particularly Hugh, Julia, David, William, Rhiannon, Cameron, Ruby, Liz, Philip, Benjamin, Ellen, Howel, and Sheila.



# Table of Contents

Abstract	3
Acknowledgements	5
Chapter 1 Introduction: Iconicity and its Place in Language	21
1.1 What is Iconicity?	21
1.2 Iconicity and Natural Language	26
1.3 Outline of the Thesis	44
Chapter 2 Outline of Sound-Shape Iconicity	47
2.1 Introduction to Sound-Shape Iconicity	47
2.2 The Phonetic Basis of the Effect	53
2.3 Experiment	66
2.4 Results	69
2.5 Discussion	89
Chapter 3 Sound-Shape Iconicity and Word Learning	99
3.1 Introduction: Cross-Situational Learning	99
3.2 Experiment 3.1: Replication of Monaghan et al. (2012)	106
3.3 Experiment 3.2: Round vs. Spiky Conditions	117

3.4	Experiment 3.3: Round vs. Spiky Conditions Replication 1	126
3.5	Experiment 3.4: Round vs. Spiky Conditions Replication 2	132
3.6	Brief Introduction to Bayesian Statistics	138
3.7	Omnibus Bayesian Statistics	143
3.8	Discussion	153
Chapter 4	Iconicity in Language Evolution	157
4.1	Introduction: Iterated Learning and the Cultural Evolution of Language	157
4.2	Experiment 4.1: Iterated Learning, Text	174
4.3	Experiment 4.2: Iterated Learning, Speech	189
4.4	Experiment 4.3: Spontaneous Word Generation	195
4.5	Discussion	199
Chapter 5	Comprehension of Iconic Gesture in Language Acquisition	207
5.1	Introduction: Child Iconic Gesture	207
5.2	Method	223
5.3	Results	230
5.4	Discussion	238
Chapter 6	Conclusions and General Discussion	241

6.1	Retrospective of the Thesis	241
6.2	Conclusions and Future Directions	246
	References	253
	Appendices	283
2.1	Ratings Scale and Instructions	283
2.2	Complete Table of Syllable Scores	284
2.3	Standard Deviation of Syllable Scores	286
2.4	Syllable Scores Minus Vowel Means	288
2.5	Syllable Scores Minus Consonant Means	290
2.6	Specification for the Stops Mixed Model	293
2.7	Specification for the Fricatives Mixed Model	293
3.1	Ratings Scale for LetterScore Norming	293
3.2	LetterScore Syllables and their Ratings	293
3.3	Specification for Mixed Model for Experiment 3.1	296
3.4	Specification for Mixed Model for Experiment 3.2	296
3.5	Specification for Mixed Model for Experiment 3.3	296
3.6	Specification for Mixed Model for Experiment 3.4	297

3.7	Specification for Bayesian Model	297
3.8	95% Highest Density Intervals for Bayesian Models for Chapter 3	298
4.1	Ratings Scale for WordScore	301
4.2	Model Specifications for Experiments 4.1 and 4.2	301
4.3	Syllable Analysis for Experiment 4.1	304
4.4	Monolingual Anglophone Analysis for Experiment 4.3	305
4.5	Syllable Analysis for Experiment 4.3	305
5.1	Stills from Object Videos from Experiment 5	306
5.2	Stills from Gesture Videos from Experiment 5	310
5.3	Trial Order from Experiment 5	313
5.4	Analysis by Items for Experiment 5	315

## List of Figures

Figure 2.1	The Articulatory System_____	55
Figure 2.2	English Vowels_____	57
Figure 3.1	Stimuli for Experiment 3.1_____	107
Figure 3.2	An Example Experiment Trial_____	110
Figure 3.3	Graph of Predicted Accuracy against Congruence by Block for Experiment 3.1_____	113
Figure 3.4	Graph of Predicted Accuracy against Congruence by Block for the +Different Model for Experiment 3.1_____	114
Figure 3.5	Graph of Predicted Accuracy against Congruence by Block for the -Different Model for Experiment 3.1_____	114
Figure 3.6	Graph of Predicted Accuracy against Congruence by Block for Experiment 3.2_____	124
Figure 3.7	Graph of Predicted Accuracy against Congruence by Block for the Rounded Condition in Experiment 3.2_____	124
Figure 3.8	Graph of Predicted Accuracy against Congruence by Block for the Spiky Condition in Experiment 3.2_____	125
Figure 3.9	Graph of Predicted Accuracy against Congruence by Block for Experiment 3.3_____	130

Figure 3.10	Graph of Predicted Accuracy against Congruence by Block for the Rounded Condition in Experiment 3.3	130
Figure 3.11	Graph of Predicted Accuracy against Congruence by Block for the Spiky Condition in Experiment 3.3	131
Figure 3.12	Graph of Predicted Accuracy against Congruence by Block for Experiment 3.4	135
Figure 3.13	Graph of Predicted Accuracy against Congruence by Block for the Rounded Condition in Experiment 3.4	136
Figure 3.14	Graph of Predicted Accuracy against Congruence by Block for the Spiky Condition in Experiment 3.4	136
Figure 3.15	Posterior Distribution for the Interaction between Condition and Congruence in Bayesian Model 1	147
Figure 3.16	Posterior Distribution for the Interaction between Condition, Category of Foil, and Congruence in Bayesian Model 1	149
Figure 3.17	Posterior Distribution for the Interaction between Condition and Congruence in Bayesian Model 2	151
Figure 3.18	Posterior Distribution for the Interaction between Condition, Category of Foil, and Congruence in Bayesian Model 2	152
Figure 4.1	Stimuli for Experiment 4.1 and 4.2	176
Figure 4.2	Graph of Transmission Error in Experiment 4.1	180

Figure 4.3	Graph of the Interaction of Shape and Generation in Experiment 4.1, Measured by LetterScore_____	184
Figure 4.4	Graph of the Interaction of Shape and Generation in Experiment 4.1, Measured by WordScore_____	186
Figure 4.5	Graph of the Interaction of Motion and Generation in Experiment 4.1, Measured by Length in Letters_____	187
Figure 4.6	Graph of the Interaction of Shape and Generation in Experiment 4.2, Measured by WordScore_____	191
Figure 4.7	Graph of the Interaction of Motion and Generation in Experiment 4.2, Measured by Length in Syllables_____	192
Figure 4.8	Stimuli Used in Experiment 4.3_____	196
Figure 4.9	Graph of the Influence of Shape on LetterScore in Experiment 4.3_____	198
Figure 4.10	Graph of the Influence of Motion on Length in Letters in Experiment 4.3_____	199
Figure 5.1	A Still from one of the Object Videos from Experiment 5__	224
Figure 5.2	A Still from one of the Gesture Videos from Experiment 5__	226
Figure 5.3	A Still from the Crucial Stage of a Trial from Experiment 5__	229

Figure 5.4 A Screenshot from Video Capture of a Subject in Experiment 5  
\_\_\_\_\_230

Figure 5.5 Graph of Frame-by-Frame Comparison between Probability of  
Subjects Looking at the Target vs. the Foil in the Gesture and in the Control  
Condition in Experiment 5 \_\_\_\_\_236

Figure 5.6 Graph of Frame-by-Frame Predicted Probability of Subjects  
Looking at the Target vs. the Foil in the Gesture Condition in Experiment 5  
\_\_\_\_\_237

Figure 5.7 Graph of Frame-by-Frame Predicted Probability of Subjects  
Looking at the Target vs. the Foil in the Control Condition in Experiment 5  
\_\_\_\_\_237

## List of Tables

Table 2.1	Previously Reported Relationships between Phonetics and Shape	50
Table 2.2	Predictions of Theories of the Mechanism of Sound-Shape Iconicity	53
Table 2.3	English Consonants	58
Table 2.4	Mean Syllable Scores by Consonant	70
Table 2.5	Mean Syllable Scores by Vowel	72
Table 2.6	The Sonorants-are-Round Generalisation is Confirmed	75
Table 2.7	The Plosives-are-Spiky Generalisation is Confirmed	76
Table 2.8	The Back Vowels-are-Round Generalisation is Confirmed	77
Table 2.9	The Rounded Vowels-are-Round Generalisation is Confirmed	78
Table 2.10	The Low Vowels-are-Round Generalisation is Falsified	80

Table 2.11	Some (but not all) Predictions of the Orthography Account are Supported_____	82
Table 2.12	Inventory of English Stops_____	83
Table 2.13	Inventory of English Fricatives_____	83
Table 2.14	Contrast Codes for the Stop Model_____	84
Table 2.15	Some (but not all) Predictions of the Lipshape Account are Supported_____	85
Table 2.16	Contrast Codes for the Fricative Model_____	86
Table 2.17	The Influence of Voicing on Obstruents Depends on the Obstruent _____	88
Table 3.1	WordScores for Stimuli in Experiment 3.1_____	108
Table 3.2	Contrast Codes for Experiment 3.1_____	112
Table 3.3	WordScores for Stimuli in Experiment 3.2_____	120
Table 3.4	Relative Congruence in Experiments 3.2-3.4_____	122
Table 3.5	WordScores for Stimuli in Experiment 3.3_____	128
Table 3.6	WordScores for Stimuli in Experiment 3.4_____	133
Table 4.1	List of LetterScores_____	183
Table 4.2	The Initial Language for Experiment 2_____	190

Table 5.1 List of Object Videos in Experiment 5 \_\_\_\_\_ 225

Table 5.2 List of Gesture Videos in Experiment 5 \_\_\_\_\_ 227



# Introduction: Iconicity and its Place in Language

The typical assumption in the study of language during the 20th Century was that the form of words has nothing to do with their meaning (de Saussure, 1916; Hockett, 1960; Levelt, Roelofs, & Meyer, 1999). However, there is evidence of non-arbitrariness ('iconicity') in the lexica of the world's natural languages, and, for spoken languages, of iconicity in the paralinguistic channels of gesture and prosody. Moreover, there is evidence that this iconicity has real psychological payoffs in the form of increased learnability and facilitated processing. This introduction will begin by defining iconicity. It will then give an overview of what we know about the role iconicity plays in natural language, and what we don't. Finally it will set out a roadmap for the rest of the thesis.

## What is Iconicity?

Though the concept of a non-arbitrary language is much older (Plato, 360 B.C.; Locke, 1690), the term *iconicity* was first introduced by the late 19<sup>th</sup> and early 20<sup>th</sup> Century American philosopher Charles Sanders Peirce in 1867. His taxonomy of signs (which is general, not merely linguistic) will be a useful place to start the explanation of what I will mean by iconicity in this thesis. Peirce's semiotics are complex, somewhat inconsistent across his long career, and rather idiosyncratic. However, roughly speaking Peirce says that one important way of classifying signs groups them into three types (Peirce, 1931-1935, 2.228; 2.229):

- Symbols: signs that are connected to their objects by convention or interpretative habit (e.g. the numeral “7” being used for sets of cardinality seven – nothing about the sign is a particular fit for its object)
- Indices: signs that receive their meaning by virtue of some real connection (causal or otherwise) to their object (e.g. smoke for fire, mercury level in a thermometer for temperature)
- Icons: signs that are connected to their object by resemblance or imitation (e.g. the word *splash*, the British Sign Language sign for CAT, a No Smoking sign, a graph visually reproducing the relations between aspects of its object)

Iconicity is the property of being an icon in the above sense. Peirce explicitly rules into the possibilities of the icon not only simple resemblances (e.g. size, shape), but also structural and metaphorical similarities of relation or structure, so right from the outset we have more than one possible kind of iconicity (a point we will return to shortly).

Before we go into the details of iconicity in natural language, I want to acknowledge some apparently problematic features of the notion of iconicity, and identify which really are problematic for us and which are not. As a *metaphysical* stance on the relationship of signs to objects, the idea of iconicity immediately draws its defenders into other questions. For instance, iconicity is defined in terms of resemblance. Any proponent of iconicity therefore has to take a stance on resemblance that doesn't end up with the result that everything resembles everything else equally (or not at all), as this would mean iconicity cannot possibly link sign to object. Establishing this means diving into debates like whether the correct account of universals is realist or a nominalist, and what the details of that account should be. Needless to say this is a thorny issue.

Fortunately this need not concern us here, as our interest is not metaphysical but psychological. As long as human minds represent relations such as resemblance, it doesn't matter whether those relations actually exist in a metaphysical sense: they can still play a role in establishing our systems of signs. Thus if the philosophical issues above trouble us, we can read 'iconicity' as meaning *iconicity relative to human mental representations*. However, the psychological view raises its own set of questions, which should be borne in mind through what follows.

Firstly: are all forms of iconicity represented in the same way? Almost certainly it's naïve to imagine that there is an iconicity module localised somewhere in the neocortex that lights up if and only if we're dealing with iconicity. Instances of iconic signs are as diverse as:

- Onomatopoeic words which directly imitate the sound of their referent
- Sound-symbolic words which e.g. exhibit structural isomorphism with their referents by using syllable reduplication to denote repetition
- Sound-shape iconicity (i.e. applying words like 'bouba' to round things and 'kiki' to spiky things': see below and Chapter 2), which appears to involve difficult-to-make-sense-of cross-modal resemblance
- Mimetic gestures or signs in sign language which schematically represent the shape of referents

Thus iconicity appears unimodally in different modalities, cross-modally, and at different levels of abstraction. It is by no means a foregone conclusion that all these different types of iconicity are represented in the same way using the same cognitive systems (Perniss & Vigliocco, 2014). Indeed later in the thesis we will see that the literature suggests that some: e.g. sound-shape iconicity, are mastered at a younger

age than others: e.g. iconic gesture, suggesting that at least some of the mechanisms involved in each case are different.

Secondly, how many different abilities are involved in any given act of understanding or producing an iconic sign? Logically, it seems that interpreting an iconic sign is an act with at least two subcomponents: recognising the resemblance between the sign and its object, and recognising that this resemblance is used *referentially*, i.e. as part of an intentional (in both senses) act of communication. Both this point and the previous one about types of iconicity can be summarised as a caveat that iconicity, though a useful category, may be cognitively heterogeneous. At present our understanding of the cognitive bases of iconicity is quite limited, but we should be ready to revise our definitions and taxonomies in order to carve cognition at the joints. However, I sketch a possible taxonomy below.

This brings us to a final question: how much a psychological understanding of iconicity will diverge from Peirce's (descriptive rather than explicatory) logical criteria as set down above. Is it the case that everything a psychologist should wish to term iconicity meets Peirce's strict definition? For instance, take the sound-shape iconicity mentioned above. It is difficult to make sense of the claim that there could be a literal resemblance between a sound and a shape. And yet most people, when confronted with the right examples, feel that there is. A number of the most plausible accounts of the effect explain it on the basis of an experientially learned association between certain speech sound (or their acoustic properties) and certain objects of given shapes (either mouth shape, or letter shape, or the size/shape of a wider range of objects in the world). In a strict Peircean sense, this is indexicality (a sign-object link by virtue of real connection) rather than iconicity (a sign-object link by virtue of real resemblance). Nonetheless, there are good reasons to think that people really do

perceive those words and those shapes as resembling one another. Therefore I think it is reasonable to cast the net wide when defining iconicity for psychological purposes: any pairing of word and referent that people perceive as resembling one another should count.

However, one thing that I explicitly exclude from my definition of iconicity is the interesting phenomenon that Monaghan, Shillcock, Christiansen, and Kirby (2014) term *systematicity*; see also Cuskley and Kirby's (2013) sensory versus conventional sound-symbolism, and Gasser, Sethuraman, and Hockema's (2010) distinction between absolute and relative iconicity (the latter in each case basically corresponding to systematicity). This is a scenario where there is *correlation* between form and meaning. A good example of this is phonestheme clusters, e.g. the English set *glow, gleam, glimmer, glister, glisten, glare*, all of which have meanings related to the emission of light<sup>1</sup>. Clearly the form of these words is not merely arbitrary, but there is seemingly no sense in which each individual sign resembles its object. For clarity, it is important to make this distinction clear at the outset, as some accounts conflate systematicity and iconicity proper (perhaps because both are present in phenomena like sound-symbolism).

One last point to avoid later confusion before we turn to natural language: iconicity and conventionality are not mutually exclusive. This is made clear by considering the onomatopoeic words for the cock's crow in different languages: *cock a doodle doo* (English), *chicchirichi* (Italian), *kikeriki* (German). Clearly each of these words is iconic, but they are also different, and conventionalised (having a lexical entry, and

---

<sup>1</sup> More generally, morphology could also be analysed as an example of systematicity at the level of the wordform (e.g. English words ending in *-able* are adjectives pertaining to some entity's propensity to undergo some process).

respecting the phonotactics of the language in question). Thus when I claim that a word is iconic, this should not be taken as a further claim that the word is not also conventionalised.

## Iconicity in Natural Language

The assumption that any sign can be matched up with any meaning has sometimes been taken as a truism (de Saussure, 1916). After all, how else could different languages have different words for the same thing ('dog', 'chien', 'perro', 'hund' etc.)? The arguments against natural connections between form and meaning have intuitive force (see Locke, 1690): signs are always partly ambiguous and their meaning can't be guessed from form alone without context (Wittgenstein, 1953). But though language is often arbitrary (sometimes necessarily so), many modes of human communication - including spoken language, signed language, gesture, and facial expression - show extensive iconicity (Perniss, Thompson, & Vigliocco, 2010; gesture will be reviewed more extensively in Chapter 5).

### **Spoken Language**

My focus over most of the rest of the thesis is on vocabulary, and I review the literature on this shortly. However it should be noted there is also substantial iconicity at higher levels such as narrative and syntax, including linear word order (Berlin, 1994; Croft, 1990, 2003; Givón, 1985, 1991; Greenberg, 1963; Haiman, 1980, 1985; Levinson, 2000; Newmeyer, 1992). For example in 'John braked and made a left turn' the implication is that the braking and the turning left came in that order (even if

the truth conditions of the sentence do not entail this). Causality can be similarly encoded in word order: in a sentence like ‘You give me a pay rise or I quit’, the implication goes beyond the literal meaning of the logical disjunction ‘or’ to imply that failing to give a pay rise will *cause* me to quit (an implication not present if the order is reversed). As causes must precede effects, this word order iconically encodes that aspect of causality. In the phrase ‘over and over and over again’ the repetition of *over* iconically enacts repetition of an event. Even constituency structure itself could be regarded as iconically encoding contiguous clusters of properties as constituents: in ‘The quick brown fox jumps over the lazy dog’, the words *quick brown fox* are grouped together in a constituent, just as the sets they refer to intersect in a single entity. Narrative is also extensively iconic, tending to unfold events in chronological order, and localise events in a particular spatial location in the same passage (unless self-consciously departing from these principles for deliberate effect).

English and other Indo-European languages are relatively poor in lexical iconicity, to which we turn now. As these are the languages in which the majority of cognitive scientists and theoretical linguists work, scholars have tended to assume that all languages are equally impoverished. Nonetheless, pockets of iconicity exist even here in the form of onomatopoeia: words that directly mimic sound (e.g. ‘bang’, ‘hiss’, ‘splash’, ‘gurgle’, ‘cock-a-doodle-doo’, ‘whack’). In many other spoken languages however, iconicity is relatively abundant – including the great majority of sub-Saharan African languages (Childs, 1994); some Australian Aboriginal languages (Alpher, 2001; McGregor, 2001; Schultze-Berndt, 2001); Southeast Asian languages (Diffloth, 1972; Watson, 2001); indigenous languages of South America (Nuckolls, 1996); Balto-Finnic languages (Mikone, 2001); and Japanese (e.g., Kita, 1997; Hamano, 1998).

Iconicity in these languages takes various forms. Japanese has a class of mimetic or sound-symbolic words that show systematic/near-systematic relationships between aspects of sound and meaning (e.g. initial consonant and size of referent), with one dictionary of Japanese mimetics listing over 1,700 entries (Atoda & Hoshino, 1995). In Japanese, reduplication of syllables often refers to repeated actions (e.g. 'goro' means a heavy object rolling, 'gorogoro' means a heavy object rolling repeatedly; 'koro' means a light object rolling, 'korokoro' means a light object rolling repeatedly). The same syllabic reduplication is seen in Siwu (a Niger-Congo language), which additionally maps unitary events onto single syllables, and signifies unitary but durative events by a lengthened vowel (Dingemanse, 2011). Sound-symbolic iconic words are used extensively in everyday conversation, and are particularly favoured in story telling as a way of bringing tales to life. In Japanese they can be found in everything from comic books to novels by Nobel-Prize winners (Schourup, 1993). Many sound-symbolic words can be described as *ideophones*, a rather loosely defined class of sound-symbolics that depict sensory imagery, not necessarily in the auditory modality (Dingemanse, 2012).

Sound-symbolism is partly conventionalised within languages. However it is not entirely arbitrary. When speakers of entirely different languages are given words in each other's languages and asked to choose the correct meaning from pairs of antonyms, they perform above chance. This holds for Japanese ideophones and English speakers (Imai, Kita, Nagumo, & Okada, 2008; Iwasaki, Vinson, & Vigliocco, 2007a; Oda, 2000; though for a failure to replicate for ideophones expressing aesthetic judgments see Iwasaki, Vinson, & Vigliocco, 2007b), and English, Czech, Hindi, and Chinese (Brown, Black, & Horowitz, 1955). Thus sound alone can be informative about meaning.

In a related phenomenon, the same associations between word-sound and shape are respected among highly geographically, culturally, and linguistically diverse populations. Back vowels and higher sonority consonants evoke large, heavy, slow, rounded things, whereas front vowels and lower sonority consonants evoke small, light, quick, jagged things (Ramachandran & Hubbard, 2001; there is evidence that other speech sound-property correlations obtain for properties like taste, colour, and brightness – see Dingemanse and Lockwood, 2015, for a review). Sound-shape iconicity will be discussed extensively in Chapter 2, so I will not elaborate here, but it is interesting to note that it appears to inform sound-symbolism: the [g] of the aforementioned *goro* (heavy object rolling, in Japanese) is a rounder, heavier consonant than the [k] of *koro* (light object rolling).

Related forms of iconic mapping dealing with non-sonic properties other than shape are also incorporated even into lexica that are not sound-symbolic. Indexicals for short distances tend to contain front vowels whereas those for big distances tend to contain back vowels (Tanz, 1971), and diminutives tend to contain high front vowels (Ultan, 1978). It is even the case that words for the mouth tend to contain bilabial consonants, whereas those for the nose contain nasals (Urban, 2011).

Iconicity is also present in prosody, particularly that of infant directed speech (Perniss, Thompson, & Vigliocco, 2010). Imagine a parent saying “the big, scary bear!” in a gruff voice, or an exasperated commuter complaining “It was soooo slooowwwwww”. Such prosodic iconicity has the potential to facilitate communication (Nygaard, Herold, & Namy, 2009; Shintel, Nusbaum, & Okrent, 2006).

Adult face-to-face communication is also extremely rich in gestural iconicity, with 30% of gestures produced in such contexts being iconic (McNeill, 1992). Such

gesture is automatically integrated with the speech channel (Chu & Kita, 2008; Kelly, Özyürek, & Maris, 2010).

## **Sign Language**

Sign language is not our primary focus here. However the visual modality – utilising hand shape, position, and movement – affords richer iconicity than is possible for spoken languages (see Perniss, Thompson, & Vigliocco, 2010 for a full review), and so provides an interesting test case. Iconicity may represent e.g. appearance (as in British Sign Language CRY, involving the outlining of tear tracks on the cheek) or motion (as in BSL AEROPLANE, which represents the motion of an airborne plane). Iconicity for action is also common (e.g. BSL HAMMER, mimetic of a hammer hold and hammering motion).

## **Human Sensitivity to Iconicity**

### **Spoken Language**

Beginning with segmental iconicity, Cuskley (2013) showed that participants are sensitive to motion iconicity, a form of iconicity common in sound-symbolic languages (as in the *goro* vs. *gorogoro* example above), and which we will revisit in Chapter 4. When asked to adjust the speed of an animation of a moving ball to match a premade nonword, participants chose to pair back vowels with slow speed, and consonant reduplication with faster speeds. The literature on sound-shape iconicity contains many instances of people's sensitivity to that form of iconicity (as

well as its benefits to word learning and processing: see Chapter 2 for a fuller discussion, or Dingemanse & Lockwood, 2015, for a review).

Adults are also sensitive to iconic prosodic contours. Nygaard, Herold, and Namy (2009), played participants novel words and asked them to guess the meaning of those words by choosing between pairs of pictures depicting antonyms on a certain dimension (e.g. hot-cold). When the words were recorded with infant-directed-speech-like prosody by people who knew their meanings, participants were able to use this prosodic information to correctly guess the meaning. However when names were recorded with the kind of prosody associated with an irrelevant dimension, prosody was no help. This shows that prosody was not simply communicating valence across all dimensions, but was rather imparting something more semantically specific. Mitterer, Schuerman, Reinisch, Tufvesson, and Dingemanse (2012) show that resynthesized versions of ideophones from five languages in five semantic domains are semantically transparent to speakers of other languages, but only if characteristic prosody is included as well as phonemes.

Shintel, Nusbaum, and Okrent (2006) examined analogue properties of the speech channel (so called 'spoken gesture' – essentially aspects of prosody) in participants who had been asked to describe the direction of a moving dot. Participants spontaneously used higher vocal pitch for an upwards-bound dot than for one moving downwards, iconically mirroring the dots' motion with their voices. Likewise, when participants had to describe dots' direction of horizontal motion they used rate of speech to iconically encode information about the dots' speed, speaking faster for faster moving dots. Similarly, Walker et al. (2010) showed that infants as young as 3-4 months are sensitive to cross-modal iconic mappings. Infants looked longer at videos of bouncing balls when their up-down motion was accompanied by a slide-

whistle sound with congruent rather than incongruent pitch change (i.e. high pitches matching high locations and low pitches matching low location).

There is also substantial evidence that from quite an early age (i.e. c. 26 months) human beings are sensitive to iconic gestures. See Chapter 5 for a full review, alternatively Özçalışkan, Genter, & Goldin-Meadow (2014). However, evidence for children younger than 26 months is extremely sparse, a problem I begin to address in Chapter 5.

## **Sign Language**

Users of sign language are aware of iconic properties of signs. Emmorey (2014) discusses iconicity in sign language extensively, showing that it affects phenomena such as metaphor and anaphora. The iconicity in sign language is sometimes so overt that awareness is taken for granted, thus many studies that show awareness really focus on processing. I will discuss some of these below. However it would be interesting to see more explicit study of the extent to which sign iconicity enters conscious awareness during production and comprehension.

## **Iconicity in Vocabulary Acquisition**

### **Spoken Language**

Imai, Kita, Nagumo, & Okada (2008) showed that word learning in children can benefit from iconicity. They created novel verbs, some of which were sound-symbolic of particular actions, others of which were not. The verbs were used as the materials

in a learning task for 3-year-old Japanese-speaking children. The children learned iconic verbs better than non-iconic alternatives. Kantartzis, Imai, and Kita (2011) found the same results for English speaking three-year-olds, indicating the benefit to Japanese speakers at least partly reflects language-general biases rather than internalisation of language-specific conventions. Equivalent results were obtained in a very similar study by Yoshida (2012, Experiment 2), for both Japanese and English speaking children.

Adult experimental studies supporting the possibility that sound-symbolism could benefit vocabulary acquisition include Monaghan, Mattock, and Walker (2012), who show that adult Anglophones are more effective at learning iconically congruent names for shapes in cross-situational learning. Lockwood, Dingemanse, and Hagoort (2016) show similar findings for Dutch adults and Japanese ideophones, using Dutch concept names rather than shapes, and a more explicit teaching method. Nielsen and Rendall (2012), show that adult Anglophones, trained on pairings drawn from a set that pairs iconically congruent names and shape, could generalise this sound-symbolism to identify further correct pairings from the same set, whereas participants trained and tested on an iconically incongruent set could not.

Observational research shows that in the course of normal learning, Japanese-speaking children acquire iconic words early (Maeda and Maeda, 1983). In keeping with this, both Maguire et al. (2010) and Saji and Imai (2013) find that Japanese-speaking caregivers use more sound-symbolic and onomatopoeic words when speaking to their toddlers than when addressing adults. Yoshida (2012, Experiment 1), in a study that had parents demonstrate how to play with toys to their children in the lab, showed that Japanese parents almost ubiquitously employed sound-

symbolism (with over a third of English speaking parents also employing vocal sound effects or onomatopoeia, and many also using a verb like 'sprinkle' that raters subsequently deemed iconic).

Perry, Perlman, and Lupyan (2015), analysing English and Spanish, two languages not generally considered highly iconic, gave speakers early acquired verbs to rate for iconicity. They found that adjectives tend to be more iconic than nouns and verbs (in keeping with previous claims about iconicity across languages: Dingemans, 2012; Imai & Kita, 2014) and that English verbs tend to be more iconic than Spanish verbs due to Spanish verbs tending to carry path information, and English verbs tending to carry manner information. Crucially, they also found that there is a negative correlation between iconicity and age of acquisition: i.e. the earlier acquired the word, the greater its iconicity was likely to be. This suggests that even in Indo-European languages that appear poor in iconicity, it may play an important role in acquisition.

Aside from this interesting study, which is limited to establishing correlation rather than causation, very little work has addressed the question of how much of a role segmental iconicity plays in lexical acquisition of non-sound-symbolic languages, or how much of a role prosodic iconicity plays in any acquisition of any language. Moreover, while gestural iconicity is somewhat better studied, and has been shown to be capable of teaching children aspects of word meaning, at least in the short term (see Chapter 5 for a review), almost all studies look at children who are already old enough to be well into vocabulary acquisition. Chapter 5 of this thesis however looks at 18-month-old English speaking children and gestural iconicity.

## Sign Language

Contrary to earlier studies (Folven & Bonvillian, 1991; Orlansky & Bonvillian, 1984) Thompson, Vinson, Woll, & Vigliocco (2012) showed that iconicity of signs (as operationalized by ratings from native adult signers) predicted sign production and comprehension by deaf infants and toddlers as reported in the BSL communicative development inventory. Moreover Perniss, Lu, Morgan, & Vigliocco (under review) show that deaf mothers of young children accentuate iconicity in their signing when referring to absent objects. However, the role of iconicity in the acquisition of these highly iconic languages remains underexplored.

## Summary of Acquisition Literature

In general, the nature of the advantage iconicity confers in vocabulary acquisition is not entirely clear: it could help make wordforms more memorable, or it could help solve what is widely described as Quine's (1960) *gavagai* problem<sup>2</sup> of referential

---

<sup>2</sup> The situation Quine outlines is this: you are with a local in a country whose language you do not speak. A rabbit hops past, at which your companion says "Gavagai"! It would be natural to translate this as something like "Lo, a rabbit!", but a moment's reflection reveals that nothing we know about the scene is inconsistent with it meaning "Stage in the life history of a rabbit!", or "Lo, undetached rabbit part!", or "Let's go hunting!".

This example is widely used as an illustration of the point that a child learning the vocabulary of its native tongue is faced with a problem that also dogs any scientist constructing a hypothesis: the underdetermination of theory by evidence. There are many interpretations consistent with what we know about *gavagai*, and with what a language-acquiring infant might know about *gavagai* too. Anything that narrows them down is therefore helpful.

However, just so that this may be acknowledged in some corner of the psychological literature, however obscure, I want to point out that what Quine intended to demonstrate with this example went well beyond the familiar underdetermination of theory by evidence: he attempted to show that the correct translation of *gavagai* is not merely underdetermined, but *indeterminate*. He wanted to show that there is no single fact of the matter about what the speaker meant: any interpretation that can be made to fit with the facts of his behaviour is equally valid.

disambiguation, or both. Almost all studies so far content themselves with trying to establish that iconicity makes a difference rather than establishing precisely what that difference is. Chapters 3 and 5 of this thesis both attempt to shed some light on the specific advantage iconicity confers on word learning.

## Lexical Iconicity in Processing

### Spoken Language

Westbury (2005) found evidence for a benefit of kiki-bouba style sound symbolism to processing in a special lexical decision task. The words used featured one of two classes of consonant, and each word was presented in a frame (a white shape on a black background) that could either be round or spiky. Westbury found that for nonwords, responses were faster when what he calls 'continuants' – actually sonorants like nasals or approximants (e.g. /m/ or /l/) – appeared in a rounded frame, and stops (e.g. /p/ or /k/) appeared in a spiky frame. In a single character letter/non-letter decision control condition he showed that this could not simply be attributed to letter shape. However, though the work is not published, other scholars have failed to replicate the first part of this experiment (Julio Santiago, personal communication).

Kovic, Plunkett, and Westermann (2010), taught participants names for pictures via an implicit learning categorisation task. The pictures were of animals, all of whose prominent anatomical features were round (in one category) or spiky (in the other).

---

Needless to say I cannot do justice to Quine's thesis here, but there is a certain irony in broadly cognitivist-mentalist-internalist scholars like Bloom (2000) deploying Quine's example as a cornerstone of their accounts of word learning; Quine's psychological proclivities were thoroughly behaviourist. See Searle (2002) for an attempt to refute Quine's indeterminacy argument.

The two categories of animals had names chosen for sound-symbolic association with round and spiky shapes – e.g. ‘mot’ (round) and ‘riff’ (spiky). Participants were assigned to one of two conditions – congruent (where ‘mot’ was the name of the category of round animals, ‘riff’ of the spiky) and incongruent (where ‘riff’ was the name for the round category, ‘mot’ the spiky). In a testing phase at the end of the experiment, participants were faster to accept congruent mappings. Furthermore, EEG readings showed an early negativity (N-200) for iconically congruent mappings as compared to incongruent mappings, which the authors suggest reflects auditory-visual feature integration. Moreover the ERP signal for accepting or rejecting pairings of round animals and round names was distinct from all others, suggesting that round-round iconicity may occupy a privileged position in the effect, a point to which we return in Chapters 2, 3, and 4.

Meteyard, Stoppard, Snudden, Cappa, and Vigliocco (2015), using a range of reading, repetition, and lexical decision tasks, found that processing of iconic words in left-hemispheric Anglophone aphasics was better preserved than that of arbitrary words. They conclude that the representation of iconic words enjoys greater redundancy than that of arbitrary words, perhaps because iconicity affords additional pathways directly between phonology and semantics.

A number of neuroimaging studies have scanned participants as they perform varieties of behavioural iconicity tasks without also having a control condition where the same tasks are performed with non-iconic vocabulary, making such studies almost completely uninformative about what extra mechanisms are involved in iconicity over and above vocabulary use generally (see Lockwood & Dingemans, 2015, p. 10, for a critique of this tendency). A few recent studies have corrected this. Lockwood and Tuomainen used EEG to compare brain activity in Japanese

speakers reading Japanese sentences with ideophonic vs. non-ideophonic adverbs. They found that compared to the non-ideophonic control, brain activity during the ideophonic sentences showed a greater P2 response and late positive complex, which they interpret as reflecting multisensory integration of sound and sensory processing. Lockwood, Hagoort, and Dingemans (2016) obtained a similar result with Dutch speakers, this time finding that P3 component and late positive complex - again interpreted as reflecting multisensory integration - predicted how much individual speakers' learning benefited from sound-symbolism. Kanero, Imai, Okuda, Okada, and Matsuda (2014), in an fMRI study on Japanese speakers, found that mimetic words for shape and motion but not arbitrary words activated the right posterior superior temporal sulcus (STS), which these authors argue is a hub of multimodal integration (though results are complicated by the fact that it is also known to process biological motion, which appeared in visual stimuli for the experiment). Kanero et al. argue that the fact that extra ideophonic activation was in the STS rather than unimodal sensory areas argues against an embodied explanation of this kind of sound-symbolism, instead favouring one where both the arbitrariness and the iconicity of the word influence its neural processing. Finally Revill, Namy, De Fife, and Nygaard (2014), using fMRI with English speakers and words from sound-symbolic languages, find that the left superior parietal cortex shows increased activation for sound-symbolic compared to non-sound-symbolic words, which again they attribute to cross-modal integration.

Study of the processing of prosodic and gestural iconicity has been sparse so far – a significant gap in the literature.

## **Sign Language**

Iconicity also has consequences for processing in signed languages (see Perniss, Thompson, & Vigliocco, 2010 for a review). To give a few examples: Thompson, Vinson, and Vigliocco (2009) worked with American Sign Language signers. In a sign-picture matching task they found that iconicity accelerated response time, as long as the iconicity accentuated a feature that was salient in the picture. Controls performing the same task with English words rather than signs showed no such effect for the same pictures. In order to confirm that these results were not simply the result of conscious strategies, Thompson, Vinson, & Vigliocco (2010) had participants perform a phonological decision task in British Sign Language (BSL). Even though this did not require access to meaning, iconicity still slowed reaction time. Thompson et al. argue that this slowing represents stronger automatic activation of the semantics of iconic signs. Vinson, Thompson, Skinner, and Vigliocco (2015) examined the influence of iconicity on picture-sign matching, phonological decision, and picture naming in BSL. They found that iconicity aids comprehension across the board, and aids production for later acquired signs.

## **Summary**

In processing research there has been more of a nominal commitment to investigating mechanisms than in the acquisition research. However, with the exception of Meteyard et al. (2015) - whose study implies that iconicity builds in redundancy to the path between phonology and semantics, adding a second route in addition to the one that holds for most of the lexicon - most of the spoken language studies do not yield strong insights in this respect. Neuroimaging studies suggest

that iconic words involve multisensory integration, but this hardly surprising: it is virtually definitional of iconic words that they are entities in one modality (speech) that evoke properties in another.

Tentatively we might want to distinguish three kinds of iconicity, each of which may well involve different mechanisms:

- Iconicity based on direct perceptual resemblance (e.g. onomatopoeia)
- Iconicity based on learned cross-modal associations (perhaps including sound-shape iconicity, and mappings between sound and other physical features, such as size and brightness)
- Iconicity based on more abstract structural isomorphism (e.g. many sign language signs, mimetic and metaphorical gesture)

We might also speculate that iconicity for action (e.g. pushing) involves different mechanisms to iconicity for fixed physical properties (e.g. size). However, we still have much to learn about what different kinds of mechanisms establish the different kinds of resemblance that define iconicity, and what mechanisms make it possible for the resemblance to be interpreted as carrying communicative/referential weight. I return to this point below in discussion of theoretical frameworks for iconicity.

Thus, to summarise the preceding four sections: iconicity sits alongside arbitrariness in the world's spoken and signed languages. The existing evidence argues that speakers are sensitive to it, and that it can have facilitatory effects on language

learning and processing. However, there remains much we do not yet understand about how this works.

## A Theoretical Framework for Iconicity

As reviewed above, most of the recent psychological, linguistic, and neuroscientific literature on iconicity has been practically minded, focused on establishing iconicity's presence in natural language or demonstrating effects that show its efficacy in language acquisition and processing. However, it is important not to lose sight of bigger theoretical questions about why iconicity should matter at all.

Perniss and Vigliocco (2014) provide a useful framework, which I will adopt as a set of background assumptions for the rest of the thesis. Perniss and Vigliocco argue that iconicity is one way in which three major problems related to language might be solved:

In the evolution of language, iconicity might have been a way to bridge between reference to the immediate context and *displaced* reference (i.e. reference to things not in the here and now) due to its power to evoke objects and phenomena not actually present. This would have allowed hominin communication to go beyond the kind of immediate functional reference to perceptually present entities (e.g. predators) seen in primate signal systems, and bridge the way to true conceptual reference, which depends on mental representations of sets of entities, none of which need to be present for reference to succeed. The process may have been driven by increasing group size, and division of labour in such a way as to necessitate spatial separation.

In language ontogeny, Perniss and Vigliocco argue that iconicity might play a critical role in establishing referentiality as children's language abilities get off the ground. This shifts the problem of word learning to one where children use prior biases or statistics in the input to match arbitrary words/signs to arbitrary referents, to one where the form of the lexical item itself also provides an ever-present clue to meaning. Moreover, it means that learning about the meaning of words can potentially happen even in the absence of the word's referent, and of useful linguistic context.

Finally, iconicity may bridge between language form and meaning by providing ready-made embodiment: inasmuch as language is understood by sensory and motor activation, iconic language forms automatically provide semantic grounding through activating those systems in appropriate ways. As Meteyard, Stoppard, Snudden, Cappa, and Vigliocco (2015) argue, iconicity can therefore provide an extra route to activation of semantics, making processing more efficient and robust.

Perniss and Vigliocco's framework provides a theory of the relationship between iconicity and natural language. However, they do not claim to be giving a mechanism-level account of the phenomenon, instead recognising that iconicity comes in many forms and is likely to draw on diverse cognitive resources (e.g. perceptual categorisation for onomatopoeia, cross-modal associations for sound-shape iconicity, abstract structure mapping for iconic gesture). Therefore in addition to an account like Perniss and Vigliocco's that sets out iconicity's importance in evolution, acquisition, and processing, we need mechanism-level theories of iconicity (a possible taxonomy is sketched at the end of the previous section).

Finally, there is another theoretical problem faced by iconicity research, one which is very little acknowledged: how does iconicity get into the vocabulary? Iconicity researchers tend to allude to the fact that iconicity is beneficial to processing and acquisition when faced with this question. That may be part of the answer, but it cannot be the whole answer: we can explain features of things that are consciously designed by human beings simply by reference to the fact that they make the thing work better (Why does the boat have a keel? – To keep it stable in the water), but this won't work for natural phenomena<sup>3</sup> (Why is the pebble smooth? – #So the water can flow around it more easily). One possibility scholars occasionally raise is that iconicity in the lexicon is a remnant of a superannuated iconic protolanguage (Kita, 2008), a kind of linguistic version of the vestigial tailbone. However, I will take a different view, one that draws crucially on the role iconicity might play in language change, which is another aspect of iconicity that remains almost entirely unstudied. In Chapter 4 I will simulate language change using iterated learning (Kirby, Cornish, & Smith, 2008), a model of cultural evolution.

This thesis adds to our sparse theoretical understanding of all these aspects of iconicity, particularly to our understanding of the mechanism of sound-shape iconicity (Chapters 2-4), and of gestural iconicity (Chapter 5), of iconicity's specific role in vocabulary acquisition (Chapters 3 and 5), iconicity's role in evolution and language

---

<sup>3</sup> Darwinian explanations for natural phenomena, at least when phrased casually, often appear to have this form (Why does the gazelle have long legs? – So it can run fast and escape predators). However this is misleading, as what these explanations actually advert to is not intentional (in both senses) causation by design (as in the keel example), but a non-intentional process (natural selection) that gives design-like results. They only work given the presupposition that natural selection is in operation. Natural selection is usually the best explanation for designer-less phenomena that nonetheless appear designed, and – to preempt my later arguments – I will explain the presence of iconicity in natural language by reference to this kind of process, following similar explanations of key features of grammar by e.g. Kirby, Smith, and Brighton (2004).

change (Chapter 4), and how iconicity gets into the vocabulary in the first place (Chapter 4 again).

## Outline of the Thesis

After this introduction, in which I have introduced iconicity and showed how it plays a role in natural language vocabulary, acquisition, and processing, I will begin Chapter 2 by introducing sound-shape iconicity, a form of spoken iconicity that I deal with extensively in Chapters 3 and 4 (both as a phenomenon to be studied for its own sake and as a case study in iconicity). I will give a brief introduction to English phonetics, in order to set the scene for the study that follows. This was a norming study for a large (in fact almost exhaustive) number of consonant-vowel syllables formed using the speech sounds of English. Though the study is simple, its major advantage is giving us far wider phonetic coverage than has been attempted before, which puts the phonetic generalisations about the effect on a firm footing for the first time, and yields insights as to the basis of sound-shape iconicity, preparing the way for Chapters 3 and 4.

Chapter 3 focuses again on sound-shape iconicity, this time employing the cross-situational learning paradigm (Yu & Smith, 2007), a artificial language learning paradigm targeted at investigating vocabulary acquisition in ambiguous contexts. I begin with a near-replication of Monaghan, Mattock, and Walker (2012). Like them, I find that sound-shape iconicity improves performance in the cross-situational learning paradigm, but unlike them my data suggests that this improvement lies not in quicker learning *per se* (in the sense of more efficient use of available cross-situational statistics, or faster or more robust encoding of memory traces), but rather

in *gavagai*-style referential disambiguation (see above). I follow up this study with a set of replications of a variation on the first study, aimed at testing whether the effect is equally strong for round and spiky pairings, as might be predicted by a lip shape based account of sound-shape iconicity. One run of the experiment gives a significant asymmetry, but the others do not, so Bayesian statistics are used to test the overall picture. The conclusion is that asymmetry in this paradigm is probably small, if it exists at all.

Chapter 4 turns to experiments that involve word production. The mainstays of the chapter are two iterated learning experiments, a paradigm that is designed to investigate the dynamics of language change by having participants learn from predecessors in a chain (like the game Chinese whispers/broken telephone). The chapter begins with an introduction to the paradigm and its theoretical motivations. I then present a simple experiment where I had participants spontaneously invent text-based words for stimuli varying in shape (round vs. spiky), and duration of motion. I find that sound-shape iconicity gets built into the names, as well as iconicity for duration of motion (with longer-moving stimuli receiving longer names). Having established that biases towards iconicity are present when generating new vocabulary, I turn to the iterated learning paradigm for a more realistic model of how languages evolve over time. In both a text-based and a speech-based version of the experiment, both kinds of iconicity are seen to emerge again. I interpret this as evidence that there could be systematic pressure towards iconicity in vocabulary even if it is not obligatory, perhaps reflecting both production biases and learning advantages, and that this may explain the presence of iconicity in natural language. I also discuss what these results mean for the hypothesis that iconicity bootstrapped our ancestors' first protolanguage. I also note that these results seem to show the

round-spiky asymmetry we saw a trace of in Chapter 3, perhaps suggesting that different aspects of sound-shape iconicity are emphasised in production tasks.

Chapter 5 shifts focus from speech to gesture and from adults to children. Though spoken languages like English are comparatively poor in lexical iconicity, face-to-face communication features an abundance of iconic gesture. If iconicity is important in language acquisition, then children acquiring spoken language may make use of this gesture to learn about the meaning of words in the absence of clear referents. Against this, it has generally been assumed that children are incapable of comprehending iconic gesture before the age of 26-months. However, this assumption is based on the weak support provided by the observation that children barely *produce* iconic gesture before 26-month: very few studies directly assess comprehension, and those that do exist gave mixed results. I decided to test whether 18-month-olds could match iconic gestures to corresponding manners of motion using the most straightforward task possible: a looking time paradigm. My results were null, however this may well reflect low power.

Finally, the conclusion summarises my findings, and takes stock of the new insights provided by the preceding chapters and what they imply about the place of iconicity in spoken language.

# Chapter 2: Outline of Sound-Shape Iconicity<sup>4</sup>

## Introduction

Sound-shape iconicity (aka the ‘kiki-bouba effect’) is the cross-linguistic preference for mapping certain sounds (e.g. back vowels and high sonority consonants) to rounded objects; and others (e.g. front vowels and low sonority consonants) to jagged objects (see Lockwood & Dingemans, 2015; and Perniss, Thompson, & Vigliocco, 2010, for reviews). It has enjoyed a healthy degree of attention in recent years, and I will use it as a case study in iconicity in the chapters that follow. This chapter comprises an introduction to the literature on this form of iconicity, and an introduction to the phonetic concepts involved in the discussion of the effect, followed by a study where I had participants rate a phonetically wide-ranging set of syllables for iconicity. This is the first such study, and is important for verifying the parameters of sound-shape iconicity, and for assessing theories of its mechanism. Chapters 3 and 4 will build on what I find here.

### **The Effect**

The classic demonstration of sound-shape iconicity is an experiment where participants are given images of two 2-dimensional shapes, one round and cloud-like, the other spiky and shard-like. They are told that one of the shapes is named e.g. ‘kiki’, the other e.g. ‘bouba’ (Ramachandran & Hubbard, 2001), and asked to say

---

<sup>4</sup> Thanks to Zoë Belk for syllable recordings, and to Katrina Shum and Arzoo Mukarram for assistance with editing of recordings and construction of experiment trial orders.

which is which. Invariably, a sizable majority of respondents (70-95%) say that 'kiki' belongs to the spiky shape, and 'bouba' belongs to the round shape.

This research, building on early work by Köhler, who used 'takete' and 'baluba' (1929), and later 'takete' and 'maluma' to avoid similarity with 'balloon' (1947; later replicated by Nielsen & Rendall, 2011), has been replicated for two-year-olds (Maurer, Pathman, & Mondloch, 2006), four-month-old infants (using a looking time paradigm; Ozturk, Krehm, & Vouloumanos, 2013), Swedish speakers (Ahlner & Zlatev, 2010), Swahili speakers (Davis, 1961), and non-Indo European speaking Namibians living a non-literate, non-industrial lifestyle (Bremner, Caparos, Davidoff, De Fockert, Linnell, & Spence, 2012). The same shape-phonetics associations pertain when shapes are depicted as 3D rather than 2D (Aveyard, 2012), when real-world objects are used instead of shapes (D'Onofrio, 2014), when the methodology used is implicit learning rather than one-shot binary choice (Monaghan, Mattock, & Walker, 2012), and where the metric is rapidity of pairing rather than accuracy (Parise & Spence, 2012; Kovic, Plunkett, & Westermann, 2010). Moreover, recent EEG work shows that iconically mismatching word-shape pairings elicit an N400-like response in 11-month-old infants, as well as left-hemispheric phase synchronisation in the beta band, indicating increased processing effort in the nascent semantic network (Asano, Imai, Kita, Kitajo, Okada, & Thierry, 2015; the authors argue that infant sensitivity to iconicity bootstraps their understanding of speech sounds' referentiality).

The only population reported to be insensitive to the effect are autistic people. Ocelli, Esposito, Venuti, Arduino, & Zampini (2013) found that in a version of the classic Köhler task high-functioning people with Autism Spectrum Disorder (ASD) diagnoses showed reduced sensitivity to sound-shape iconicity, and low-functioning people with

ASD showed none at all. Ocelli et al. tentatively attribute this to poorer multisensory integration in ASD. However it is difficult to rule out the possibility that difficulty with verbal instructions (people with ASD often have difficulty with verbal information) was part of the cause of the diminished effect. Drijvers, Zaadnordijk, and Dingemanse (2015) find that Dutch dyslexics also show the effect but at diminished strength. They interpret this as reflecting difficulties in cross-modal processing in dyslexia, though it would also seem to be highly consistent with the claim that the effect partly arises from learned associations between orthography and phonology (see discussion of Cuskley, Simner, & Kirby, 2015, below).

The speech sounds identified as spiky are typically plosives (Aveyard, 2012; Monaghan, Mattock, & Walker, 2012; Nielsen & Rendall, 2011, 2013; Ramachandran and Hubbard, 2001), voiceless obstruents (Ahlnér & Zlatev, 2010; D'Onofrio, 2014), front vowels (Ahlnér & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2011), high vowels (Ahlnér & Zlatev, 2010; Monaghan et al., 2012), and unrounded vowels (Nielsen & Rendall, 2013); whereas the sounds identified as round are sonorants (Ahlnér & Zlatev, 2010; Monaghan et al., 2012; Nielsen & Rendall, 2013), voiced plosives (D'Onofrio, 2014; Ramachandran & Hubbard, 2001), labial consonants (D'Onofrio, 2014), back vowels (Ahlnér & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2013), low vowels (Ahlnér & Zlatev, 2010; Monaghan et al., 2012), and rounded vowels (Nielsen & Rendall, 2013). See Table 2.1 for a summary.

Round Sound	Source	Spiky Sound	Source
Sonorants	Ahlner & Zlatev, 2010; Monaghan et al., 2012; Nielsen & Rendall, 2013	Voiceless Obstruents	Ahlner & Zlatev, 2010; D'Onofrio, 2014
Voiced Plosives	D'Onofrio, 2014; Ramachandran & Hubbard, 2001	(Voiceless) Plosives	Aveyard, 2012; Monaghan et al., 2012; Nielsen & Rendall, 2011, 2013; Ramachandran and Hubbard, 2001
Labials	D'Onofrio, 2014		
Back Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2013	Front Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2011
Rounded Vowels	Nielsen & Rendall, 2013	Unrounded Vowels	Nielsen & Rendall, 2013
Low Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012	High Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012

*Table 2.1: Natural classes of phoneme described as round or spiky in the literature*

## Mechanisms

The mechanisms underlying sound-shape iconicity remain unclear, but there are a number of plausible suggestions, each of which makes predictions about the boundaries of the effect (summarised in Table 2.2 below). Ramachandran and Hubbard (2001) suggest that the effect is based on cross-modal analogy between visual shape and articulatory gesture, implying a non-visual representation of articulation mediating between sound and shape (p. 19). The idea is that e.g. 'sharp' sounds are somehow metaphorically linked with 'sharp' articulatory gestures. Ramachandran and Hubbard also make the alternative suggestion that 'cross-wiring' (p. 21) of auditory and visual brain maps creates a (basically arbitrary) link, with associations depending on contingencies of human brain architecture.

Another possibility is that the effect arises out of the fact that certain phonetic properties of sounds are diagnostic of physical properties of the animals that make them. It is interesting to note that some spiky sounds also tend to suggest smallness, and some round sounds suggest largeness (Dingemanse & Lockwood, 2015; Perniss, Thompson, & Vigliocco, 2010). If shape associations piggyback on size associations then those size associations may, like prosodic expression of emotion, have biological roots in the use of pitch of vocalisation as a cue to body size (a deeper pitch and deeper resonances corresponding to a larger hence more threatening animal: Ohala, 1984, 1994; Xu, Kelly, & Smillie, 2013). Use of sound-shape iconicity to signal body size may show up in e.g. sexual dimorphism in English names (Pitcher, Mesoudi, & McElligott, 2013).

An alternative mechanism would operate via unimodal mappings. Ramachandran and Hubbard (2001) note that some of the sounds that evoke roundness (e.g. rounded vowels) involve literal rounding of the lips (see also D'Onofrio, 2014; interestingly, sign languages also feature lip iconicity based on the distinction between round and thin lips: Sandler, 2009). This raises the possibility that round shapes connote round objects via a unimodal mapping between representations of lip shape and representations of object shape. The lip shape representations could be primarily visual, or primarily motoric, but would – crucially – represent actual roundness, rather than representing roundness through metaphorical associations of the sort that Ramachandran and Hubbard posit for spikiness. Unlike other accounts, this one predicts asymmetry: round sound-shape associations should be stronger than spiky ones, because round sounds involve literal rounding of an articulator, whereas spiky sounds do not involve any comparable spikiness. This dissociation is not something that the classic kiki-bouba experiment is able to test: because there

are only two words and two shapes, the determination of one (hypothetically stronger) sound-shape pairing automatically determines the other (weaker or absent) pairing. Intriguingly, what evidence there is (in the form of ERP data) suggests that the round association may be privileged over the spiky one in processing (Kovic, Plunkett, & Westermann, 2010 – though note that their paradigm was not capable of separating the associations behaviourally).

Yet another suggestion is that sound-shape iconicity is substantially driven by the forms of the letters typically used to represent phonemes. Cuskley, Simner, and Kirby (2015), reviewing the literature on the kiki-bouba effect, make a strong case that the evidence for the presence of the effect among non-literate participants (including young children) and speakers of languages that do not use Latin script is much weaker than generally supposed. The children in Maurer, Pathman, and Mondloch (2006) were old enough to have plausibly had some exposure to orthography; Ozturk, Krehm, and Vouloumanos' (2013) infant result has reportedly failed to replicate (though these data are not published); and almost all reports of the effect in adults test populations one would expect to be familiar with Latin script.

At least one robust study demonstrating the effect in a nonliterate population does exist (Bremner, Caparos, Davidoff, de Fockert, Linnell, & Spence, 2013), but this shows a weaker effect than that usually reported with literate Anglophones (82% iconic matches rather than the 95% reported by Ramachandran and Hubbard). Cuskley et al. point out that this means the existing data are consistent with the effect being substantially driven by orthography, rather than by sound as usually assumed. They present data showing that letterform is a strong predictor of sound-shape iconicity not only in the written modality, but in the spoken modality (when using consonants that are typically represented with a single canonical letter).

Theory	Round Predictions	Spiky Predictions
Cross-modal analogy	Smooth sounds – approximants?	Sharp sounds - plosives?
Cross-wiring	?	?
Lip shape	Labial/rounded: [p, b, m, w, o, u]	Everything else?
Orthography	[w, b, p, s, d, m, o, u]	[k, z, t, v, i]
Environmental statistics	Low pitches?	High pitches?

*Table 2.2: Predictions of different accounts of the mechanism of sound-shape iconicity*

## The Phonetic Basis

The experiment section and my discussion of it feature fairly extensive comment on the phonetics of English speech sounds, so it will be useful to provide a brief introduction to this topic (most of this draws from Ladefoged, 2001, which provides a much fuller introduction). I will introduce some basic ideas in phonetics/phonology, and then provide a guide to the properties of the vowels and the consonants found in English.

Phoneticians and phonologists find it useful to distinguish between phonemes and allophones. Phonemes are basic sounds of a given language<sup>5</sup> whereby the

---

<sup>5</sup> Technically, inasmuch as mainstream phonology makes use of the concept of the phoneme (rather than more abstract units like features, elements, autosegmental tiers etc.), the phoneme is deemed to be a mental representation with a default phonetic realization that can

substitution of one for another can change one word to another (e.g. *bat* vs. *pat*). Non-phonemic allophones of the phonemes are variations on the basic sounds. These can appear in their place in certain contexts, but they do not represent separate phonemes of the language in question. In what follows I will ignore allophones. For instance, in most varieties of North American English, both /d/ and /t/ are realised as the somewhat different sound [ɾ] (i.e. the alveolar tap) before unstressed vowels<sup>6</sup>. As this sound only appears in English as an allophone of other sounds, I do not include it in the study presented below. Incidentally, [ɾ] is presented in square brackets because it is the sound actually pronounced (rather than the underlying representation, written using slash brackets, hence /d/). I will continue to use square brackets when writing about speech sounds for the rest of this chapter, given that I am interested in sounds as they are actually produced rather than e.g. as they are stored in lexical entries.

The [ɾ] example also helps bring out the point English, like all languages, displays a complex system of patterns in terms of how a given sound is realised in a given speech context (this is what phonologists make a living studying). For simplicity, we will ignore this in what follows, treating each phoneme in its default form, i.e. the one it takes when produced clearly in an unmarked context.

---

be stored as part of the phonological component of morphemes' lexical entries, rather than a sound *per se* (see e.g. Hayes, 2009, for an introduction).

<sup>6</sup> Note that a sound is only phonemic or allophonic in relation to the grammar of a particular language. In Spanish, [ɾ] is a phoneme.

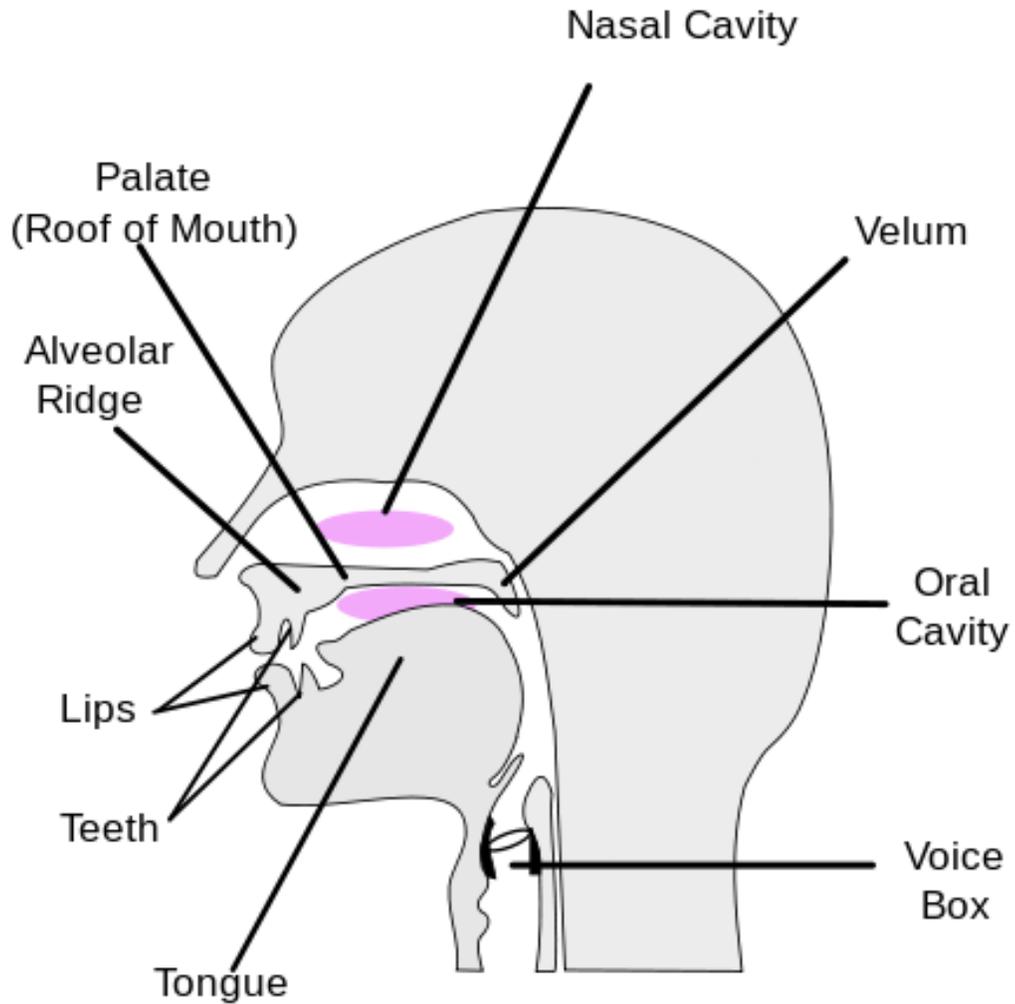


Figure 1.1: The articulatory system. N.B. the vocal folds are located in the voice box. Reproduced from Meg Smith under a Creative Commons BY-SA 3.0 licence.

Starting with vowels: the vocal folds (colloquially vocal cords or voice box – see Figure 1.1 for the location of these and other parts of the articulatory system) are membranes stretched across the larynx, which vibrate open and closed during much of speech production to create the *fundamental frequency* of speech. In English, all vowels are produced with vocal folds vibrating. Fundamental frequency carries information about stress, intonation, and the speaker’s identity and emotional state, but it is not typically used to differentiate phonemes. Instead vowels are

differentiated from each other by how the position of the tongue and the shape of the lips change the shape of the vocal tract to mould the sound coming from the vocal folds (in English all vowels' default pronunciation is with the nasal tract closed off, though there are many languages where this is not true).

Acoustically, vowel sounds are differentiated from one another by their *formants*. When the vocal folds vibrate, each opening sends a pulse of air into the vocal tract, setting it vibrating at its characteristic resonant frequencies (like tapping on a glass bottle part filled with liquid). These frequencies are the formants. Due to the complexity of the shape of the vocal tract, it has a number of formants, but the most important for vowels are the first formant, aka  $F_1$ , which is the lowest in pitch, and the second formant,  $F_2$ , which is the second lowest in pitch. The position of the tongue and lips during the production of a vowel are those that produce the vowel's characteristic formants.

At one time it was thought that  $F_1$  depended on the height of the highest point of the tongue, with greater height leading to a lower first formant, and that  $F_2$  depended on how far back the highest point of the tongue was, with more backness leading to a lower second formant. It is now known that the situation is considerably more complex than this, with tongue position having multiple dimensions, but height and backness terminology is still used in the description of vowels. High vowels have (confusingly) a low frequency  $F_1$  (e.g. c. 310 Hz for [u]), whereas low vowels have a high frequency  $F_1$  (e.g. c. 710 Hz for [ɑ:]). Back vowels have low frequency  $F_2$  (e.g. c. 870 Hz for [u]) whereas front vowels have a high frequency  $F_2$  (e.g. c. 2250 Hz for [i]). See Figure 1.2 for the position of some English vowels in acoustic space.

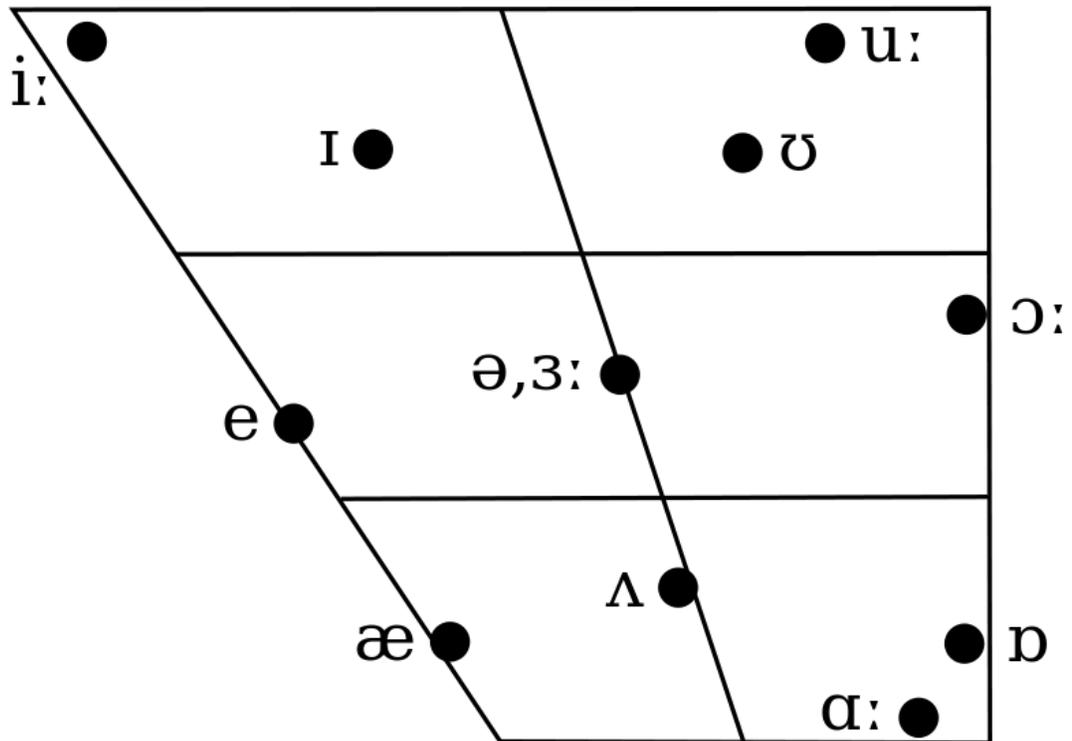


Figure 1.2: English vowels. The y-axis represents height, with more highly placed vowels having lower  $F_1$ . The x-axis represents backness, with more rightwards vowels having lower  $F_2$ . The accent shown here is Received Pronunciation, a “standard” (i.e. high-status) dialect of British English. Reproduced from Roach (2004) under a Creative Commons BY-SA 3.0 licence.

As well as tongue shape, the formants of vowels depend on rounding of the lips. To a first approximation lip rounding is acoustically realised as a lowering of  $F_2$ , meaning that rounding and backness work in tandem. Perhaps not coincidentally, all rounded vowels in English are also back vowels.

Table 2.3 depicts the consonants of (many dialects of) English:

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Glottal
<b>Plosive</b>	p b	f v		t d	ʃ ʒ			k g	
<b>Fricative</b>			θ ð	s z					h
<b>Affricate</b>					tʃ dʒ				
<b>Nasal</b>	m			n				ŋ	
<b>Lateral Approximant</b>				l					
<b>Approximant</b>	w					ɹ	j	w	

*Table 2.3: The consonants of ("standard") English. Consonant symbols are drawn from the International Phonetic Alphabet (IPA). Columns represent places of articulation, moving backwards through the vocal tract from lips to glottis. Rows represent manners of articulation, arranged top to bottom by increasing sonority (i.e. amplitude). Where two phonemes appear in the same cell, the right hand phoneme is the voiced version of the voiceless left hand phoneme. All English nasals and approximants (lateral or not) are voiced. [w] appears twice because it is deemed to have two places of articulation - it involves narrowing of the vocal tract at both the velum and the lips. Affricates are represented by double symbols because they can be regarded as the fusion of a stop and a fricative. See Table 2.3 for examples of words featuring these consonants.*

All languages tend to alternate vowels with (clusters of) consonants. The vowels are defined by fairly unimpeded flow of air through the vocal folds and the vocal tract and out of the mouth. Consonants typically involve some kind of impediment to this flow of air (see Table 2.3 for consonants of English). Often, consonants are acoustically identifiable according to the ways they influence the surrounding vowels. In English, consonants are distinguished by three properties:

- 1) Manner of articulation: how the flow of air is modified
- 2) Place of articulation: where the flow of air is modified

- 3) Phonation: the state of the vocal folds during articulation, which in English means when/whether they are vibrating

Starting with manner, plosives (also known as stops) are instances of complete closure of the vocal tract. Acoustically their hallmark is brief period of (near) silence in the speech stream. Fricatives do not involve complete closure, but do involve enough narrowing of the vocal tract to create turbulence, which manifests acoustically as high-frequency noise. Phonetically, affricates can be regarded as a stop followed by a fricative, though it should be noted that as far as the grammar is concerned they seem to be treated as a single sound produced by single gesture. Nasals (or more formally nasal stops) involve complete closure of the oral tract, but the lowering of the velum (see manner explication below) to allow air from the lungs to pass through the nose. Acoustically these are characterised by an abrupt shift from the sound produced before the nasal to a somewhat vowel-like pattern with a very low first formant, and back again. Approximants are produced by narrowing the vocal tract without closing it enough to create turbulence. The only distinction between lateral approximants and standard approximants is that in the case of laterals the opening left in the tract is to one or both sides rather than being central. In general, approximants have a vowel like qualities; indeed the glides [w] and [j] are extremely similar to the vowels [u] and [i]<sup>7</sup>.

Place of articulation is the part of the vocal tract where constriction or modification happens. This can be at any point from the vocal folds themselves right up to the end of the tract, the lips, and potentially at multiple points at the same time. A

---

<sup>7</sup> Arguably there is no set of physical, phonetic criteria that would succeed in classifying these sounds as consonants while successfully designating everything that is intuitively a vowel as a vowel. Rather the distinction between consonant and vowel must be found in where each segment is placed in the representation of syllable structure. In the interests of keeping to the point, this is another interesting issue I gloss over.

consonant typically involves an active articulator and a passive location of articulation. The active articulator is usually the tongue, the very end of which is the tip, followed by the blade, and then (moving back) the front, centre, back and root. Bilabials are formed by bring the lips together, labiodentals by bring the upper lip to the lower incisors, dentals by bringing the tip or blade of the tongue to the upper incisors, alveolars by bringing the tip or blade of the tongue to the alveolar ridge (behind the front teeth), postalveolars by bringing the blade of the tongue to the back of the alveolar ridge, retroflexes by bringing the underside of the tongue to the back of the alveolar ridge, palatals by bringing the front of the tongue to the hard palate (which is behind the alveolar ridge), velars by bringing the back of the tongue to the velum or soft palate (which is behind the hard palate), and finally glottals involve constriction of the vocal tract at the glottis, i.e. vocal folds. Acoustically, different places of articulation show up as variations on the acoustic pattern typical of a manner of articulation, particularly in the placing and transition of the formants moving out of the last vowel and into the next.

Finally, phonation is the state of the vocal folds during articulation of a consonant. Though other languages utilise other glottal states (i.e. murmured and creaky voice), the only distinction important in differentiating English phonemes is voicing versus voicelessness. Voiced consonants are produced with the vocal folds vibrating, whereas voiceless consonants are produced without. In English all approximants, nasals, and vowels are voiced, so the distinction only applies to obstruents (i.e. plosives, fricatives, and affricates). The relation between voiced and voiceless fricatives is perhaps the easiest to describe: the voiced version is basically the voiceless version plus vocal fold vibration and the low frequencies that it adds to the acoustics. Plosives are more complex: contrary to what the nomenclature seems to

suggest, English voiced plosives typically do not feature any voicing during the stop closure. Rather, they are stops where voicing resumes immediately upon the release of the stop<sup>8</sup> (i.e. at the beginning of the following vowel). Voiceless plosives are also unvoiced during stop closure, but feature *aspiration* on release – a short burst of fairly high frequency noise caused by air rushing through the newly formed opening in the vocal tract. Affricates are combinations of their component voiced or voiceless stops and fricatives.

To conclude discussion of English phonemes, a note on phonological features. Much of the phonological literature talks of segmental material (i.e. phonemes, as opposed to tones, stress, syllable structure etc.) in terms of distinctive features, typically conceived of have a value of +, -, or 0/N.A. for any given segment. Fundamentally, segmental material is represented as feature sets, and phonemes, on this view, are simply completely specified feature sets. Particular values, or sets of values, of features define *natural classes* of phonemes. For instance all phonemes articulated using the lips are the natural class [+labial], all obstruents articulated labially are the natural class [+labial, -sonorant] etc.. Historically, it has been widely assumed that phonological rules and constraints that refer to segmental material do so via features defining natural classes. This approach is driven by the fact that many interesting phonological generalizations seem to be readily expressible in these terms (Hayes, 2009). I make some use of this terminology in the remainder of the chapter as shorthand for phonetic properties, but for the most part I have not found it useful to adopt a feature-based analysis of sound-shape iconicity in any principled sense. This

---

<sup>8</sup> Other languages, such as French, do feature vocal fold vibration before stop release in [b]. The interval between stop release and voicing onset is called Voice Onset Time (VOT). VOT can be positive or negative, and varies fairly continuously between languages. Sindhi, a language of South Asia, distinguishes not two but three alveolar plosives on the basis of VOT. To the best of my knowledge there is no quantitative investigation of how VOT influences the sound-shape iconicity properties of plosives. It would be an interesting experiment.

is because the set of round and the set of spiky phonemes clearly cut across natural classes, meaning that whatever generalizations we can make to capture this phenomenon are probably not best framed in those terms. Rather than being a result of the kinds of representations assumed in phonology, it appears to me that sound-shape iconicity may be grounded in representations of acoustic and articulatory phonetics (or, others might argue, orthography), as discussed below.

### **Outstanding Questions in Sound-Shape Iconicity**

In spite of the considerable work devoted to this phenomenon in the 15 years since Ramachandran and Hubbard (2001), some rather basic questions about the effect remain unanswered.

Firstly: what are the parameters of the effect? There are certain more or less universally acknowledged phonetic contrasts that correlate with shape judgments (e.g. back vowels are round, front vowel are spiky), but I am not aware of a single study that attempts a systematic sweep of the phonetic space of a particular language. Instead previous studies presuppose these correlations, drawing a small sample of stimuli from within each natural class (e.g. Aveyard, 2012; D'Onofrio, 2014; Monaghan, Mattock, & Walker, 2012), or even eschewing a principled approach altogether and relying on intuitions about the roundness or spikiness of stimuli (e.g. Köhler, 1929; Ramachandran & Hubbard, 2001). At best this haphazard approach may mean that there are informative marginal patterns in the effect that we are not yet aware of, at worst it could mean that we are fundamentally mistaken about some of our basic generalizations, undoing all of the work that has been built on them.

Secondly: how do interactions between phonetic features and different phonemes affect the effect? Are a voiceless bilabial stop and a bilabial nasal as different in shape from one another as an alveolar stop and an alveolar nasal? Does [i] add as much spikiness to [k] as it does to [w]? To the best of my knowledge there has been no explicit investigation of these questions, but they may yield important insights into how the effect works.

Thirdly: what is/are the mechanism/s of the effect? I presented various speculations above, all of them intelligent and plausible. However, in spite of honourable exceptions like Cuskley, Simner, and Kirby (2015), surprisingly few studies appear to be designed in order to try and distinguish between different mechanisms. Given that the existence of the effect and its benefits to word learning in certain contexts are now well attested, I would argue that this is one of the more pressing issues in this line of research. Both this point and the last will be addressed to a certain extent in the remainder of this chapter.

Finally: how exactly does sound-shape iconicity enhance word learning? This is a microcosm of the wider uncertainty about how iconicity enhances word learning, and just as in the general case, few studies focus on mechanism. I will begin to explore this question in the next chapter (Chapter 3).

## **The Study**

In the remainder of this chapter I present results from a study where I had Anglophone participants rate the roundness vs. spikiness of a near-exhaustive set of the consonant-vowel syllables that can be constructed using the English phonemic inventory. Clearly this is an extremely simple design, but the advantage is that it

allows a wider sweep of the phonetic space than has been attempted before. These data will serve several purposes. Firstly, while there are widely accepted correlations between distinctive features and iconic properties, e.g. front vowels and voiceless plosives are spiky, and back vowels and approximants are round, previous studies presuppose these correlations, without thoroughly verifying them. It is important to check that the widely assumed relationships between phonetics and iconicity are in fact general across phonemes, and robust across phonemic contexts. If this can be verified, then I can build on these assumptions with confidence in the subsequent chapters involving sound-shape iconicity.

The most common claims are that higher sonority consonants are rounder than low sonority consonants (AhIner & Zlatev, 2010; Monaghan, Mattock, & Walker, 2012; Nielsen & Rendall, 2013), and that back/rounded vowels are rounder than front/unrounded vowels (AhIner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2013). I predict that these claims will turn out to be correct. I will also check less ubiquitous associations that have been suggested in the literature, namely that low vowels are rounder than high (AhIner & Zlatev, 2010; Monaghan et al., 2012), and that bilabial consonants are particularly round (D'Onofrio, 2014). The status of these predictions is less secure. In all cases though, it will be interesting to see if there are exceptions to the generalisation.

Secondly, I will use my data to address some specific and less ubiquitously accepted proposals made in the sound-shape iconicity literature. Fort, Martin, & Peperkamp (2015) found that for Francophones, consonants have a bigger effect on sound-shape iconicity than vowels (a result that could not be attributed to an onset bias, as they tested both consonant- and vowel-initial words). This result is also supported by Nielsen and Rendall (2011; though as noted in Nielsen & Rendall, 2013, both kinds

of phoneme seem to influence the effect). I predict that may data will also show this pattern.

Cuskley, Simner, and Kirby's (2015) finding that letterform predicts sound-shape iconicity could find an alternative explanation in Roman script having evolved to match pre-existing biases towards sound-shape iconicity. Such a claim would be supported if – for instance – it were shown that disparate writing systems made the same kinds of sound-shape mappings, without any such mappings being present in any common ancestor. Little systematic investigation of this exists, though Koriat and Levy (1977) show that Hebrew speakers judge Hindi and Japanese vowel orthography to have iconic properties mirroring those of their phonemic referents. Alternatively (and perhaps most plausibly), the right account of the kiki-bouba effect could be a middle ground between a story entirely based on orthography and one that ignores it completely, as pointed out by Cuskley et al.. It might go something like this: a weak form of the effect exists in the absence of any exposure to orthography, but orthography that reflects the effect (perhaps due to evolution of scripts towards iconic forms) amplifies the effect in users of that writing system. This kind of scenario would explain why the kiki-bouba effect is present among non-literate people (Bremner et al., 2013), but in a weaker form.

Clearly, definitively resolving this question will involve robust and replicable experiments on whether the effect exists among people who couldn't have picked it up from a writing system (either because their writing system doesn't make the same sound-shape mappings, or because they have no writing system), something I cannot offer here. However, I can attempt to ask whether letterform can be a complete explanation of the results I have obtained.

Finally, I will also use the dataset to begin to explore another set of questions that have been little addressed so far, but which may provide interesting insights into sound-shape iconicity: whether different aspects of a segment (i.e. manner and place of articulation) interact when determining its iconic value or whether they can be treated as additive factors, and how much the iconic contribution of a consonant depends on the neighbouring vowel (and vice versa). It is difficult to form predictions here, as this question is so little explored, but any interactions will be of interest.

## Experiment

We had participants rate a set of CV syllables nearly exhausting the combinations provided by the English phonemic inventory. Though the methodology here is extremely simple, it allows us to perform a much wider sweep of the phonetic space than any previous study, testing both previous generalisations about which classes of phonemes connote roundness and spikiness, and how much these iconic properties depend on phonetic context (i.e. which phoneme is a given phoneme's neighbour in a syllable).

## Methods

**Subjects** Were 51 native Anglophones recruited through the website Prolific Academic ( $M = 36.7 \pm 11.8$ , 18 women). Previously four participants were excluded, two for admitting to not being native English speakers, two for giving answers that indicated they had not engaged with the task (namely reaction times that were regularly less than 0.5s, or long strings of answers with the same rating). Fifty

participants took part in one or other of two sub-studies, each one featuring half of the syllable set. One participant took part in both.

**Materials** Comprised 576 consonant-vowel syllable tokens recorded by a trained North American linguist. The linguist is an Anglophone Canadian with speaking accent nearly equivalent to General American (barring phonological phenomena such as Canadian raising, which did not apply here), but adjusted her pronunciation here when recordings vowel sounds outside her accent.

Syllable forms were constructed by pairing each one of 24 consonants with each one of 24 vowels (including a number of diphthongs, some varying in length: see Tables 2.3 and 2.4 for a list of syllables). The consonants represent the set present in Received Pronunciation, General American<sup>9</sup>, and many other dialects of English. Vowel inventories are much more variable between dialects of English, and therefore the vowels do not represent the vowel set of any single accent of English, but rather a rich sample of the vowel space, each member of which is present in some significant dialect of English. This is justifiable because our aim is not to investigate the properties of some particular phonological system, but rather the properties of different parts of the phonetic space.

The resulting syllables vary as to whether they are phonotactically licit in English. English phonotactics prohibit open syllables with lax vowels in the nucleus position, and words beginning with the consonants [ŋ] (as in **si**ng) and [ʒ] (as in **treas**ure). Again, this decision is justifiable in terms of my interpretation of sound-shape iconicity being phonetic rather than phonological, and also in terms of the extra coverage of the phonetic space that this affords us.

---

<sup>9</sup> The voiceless labiovelar approximant [ɱ], corresponding to a voiceless [w], and used to distinguish *which* from *witch*, has merged with [w] in most contemporary British and North American accents, and was therefore omitted.

Post-data collection checks brought to light that minor labelling mistakes by RAs had led to the omission of ten of the 576 syllables: [ðu, ðeə, ðʊə, niə, lʊə, ʒʊə, ʌ, ɪ, ʊ, go:]. These syllables are omitted from tables below, and left out of averages and statistics. They had been replaced with duplicates of other syllable recordings; their treatment in analyses is detailed below.

## **Procedure**

Equal numbers of participants were run on four separated procedures using the online testing platform Testable. Each experimental procedure consisted of 288 trials where participants were asked to rate spoken clips on a scale of 1-7 in terms of how round or spiky they seemed. At the beginning of the procedure participants received a set of instructions, including a visual demonstration of the kinds of shape that should be taken to define the endpoints of the ratings scale (see Appendix 2.1 for an example of the instructions). It was emphasised that the study was based on sound and that participants should be in a quiet place with functioning speakers or headphones. Participants were reassured that there were no right or wrong answers, and that they should go with their instinct.

Each trial took the form of a syllable clip that automatically played at the beginning of the trial, and a written instruction of the form of e.g. “Please rate the audio clip on a scale of 1 (roundest) to 7 (spikiest)”. Participants entered their rating by pressing a keyboard key 1-7. The next trial began immediately on the participant’s response.

Procedures one and two featured one (randomly chosen) half of the syllable recordings, procedures three and four featured the other half. Procedures one and two featured the same randomly generated order of trials, differing only on which end

of the scale was assigned to roundness vs. spikiness. The same relationship held between procedures three and four<sup>10</sup>. The fact that individual syllables (though note, not individual phonemes) consistently appeared in the same contexts means we should place limited weight on individual data points, in case of context effects.

## Results

Responses with reaction times of less than 0.5s were discarded as slips of the finger. Responses with reaction times of greater than 25s were also discarded as unreliable, given that there was no way for participants to play the clip for a second time if they took a break. Thirty-two ratings were discarded this way. For participants who had received a procedure where 1 = spikiest and 7 = roundest, ratings were flipped such that 1 became 7; 2 became 6 etc.. This ensured that all data was coded such that 1 was the roundest rating, and 7 the spikiest.

## Summaries

I calculated the mean rating for each syllable, which I will refer to as that syllable's *syllable score*: these data form the basis of the results that follow. For the small number of syllables that appeared in procedures multiple times due to RA error ([ðeɪ, ði, ðɪʊ, kɔ:, li, nəʊ, ɹɑ, ʒeə] twice, and [ɹæ] three times), the mean was calculated as the mean of the means of each trial position featuring the syllable (effectively meaning that each trial position the syllable appeared in was given equal weighting in its average, regardless of any excluded trials).

---

<sup>10</sup> Ideally, each participant would have received an independent random ordering of stimuli, and would have been able to see a ratings scale during each trial. Unfortunately neither of these things were possible using Testable, which was nonetheless vastly more usable than alternative internet platforms such as Qualtrics, and much more practical than running the study in the lab.

Table 2.4 shows the roundness-spikiness ranking for consonants, and Table 2.4 shows the ranking for vowels. Appendix 2.2 shows the full set of results by syllable. Appendix 2.4 presents the standard deviation for each syllable. Appendix 2.4 shows the results for consonants relative to the mean of the vowel they were paired with, and Appendix 2.5 shows the results for vowels relative to the mean of the consonant they were paired with: these tables effectively illustrate whether a given phoneme contributes extra roundness or spikiness to its syllable, given the phoneme it is paired with.

Spikiness Ranking	Consonant	Mean Rating	Normalised Rating	Place	Manner	Example word
1	<b>k</b>	5.11	1.09	Velar	Voiceless Plosive	<b>kit</b>
2	<b>t</b>	4.73	0.71	Alveolar	Voiceless Plosive	<b>tick</b>
3	<b>tʃ</b>	4.72	0.71	Post-Alveolar	Voiceless Affricate	<b>chick</b>
4	<b>z</b>	4.65	0.63	Alveolar	Voiced Fricative	<b>zip</b>
5	<b>g</b>	4.42	0.41	Velar	Voiced Plosive	<b>gap</b>
6	<b>dʒ</b>	4.42	0.40	Post-Alveolar	Voiced Affricate	<b>jam</b>
7	<b>v</b>	4.19	0.17	Labiodental	Voiced Fricative	<b>van</b>
8	<b>p</b>	4.18	0.17	Bilabial	Voiceless Plosive	<b>pan</b>
9	<b>ʒ</b>	4.16	0.15	Post-Alveolar	Voiced Fricative	<b>leisure</b>
10	<b>d</b>	4.09	0.07	Alveolar	Voiced Plosive	<b>dock</b>
11	<b>ð</b>	4.07	0.06	Dental	Voiced Fricative	<b>thief</b>
12	<b>θ</b>	4.02	0.01	Dental	Voiceless Fricative	<b>thy</b>
13	<b>ʃ</b>	4.00	-0.01	Post-Alveolar	Voiceless Fricative	<b>sheep</b>
14	<b>s</b>	3.99	-0.02	Alveolar	Voiceless Fricative	<b>soup</b>
15	<b>f</b>	3.96	-0.06	Labiodental	Voiceless Fricative	<b>fit</b>

16	<b>b</b>	3.79	-0.23	Bilabial	Voiced Plosive	<b>bit</b>
17	<b>n</b>	3.68	-0.34	Alveolar	Nasal	<b>nip</b>
18	<b>h</b>	3.66	-0.36	Glottal	Fricative	<b>hip</b>
19	<b>ŋ</b>	3.60	-0.42	Velar	Nasal	<b>sing</b>
20	<b>ɹ</b>	3.52	-0.49	Retroflex	Approximant	<b>rap</b>
21	<b>l</b>	3.51	-0.51	Alveolar	Lateral Approximant	<b>lip</b>
22	<b>m</b>	3.43	-0.59	Bilabial	Nasal	<b>map</b>
23	<b>j</b>	3.41	-0.61	Palatal	Approximant	<b>yacht</b>
24	<b>w</b>	3.07	-0.94	Labial-Velar	Approximant	<b>wind</b>

*Table 2.4: Mean syllable scores by consonant. Rows representing syllables that are phonotactically illicit at the start of a word are highlighted in red. The normalised rating column simply shows the consonant's mean syllable score minus the grand mean: i.e. its rating relative to the average syllable. As well as place and manner information, the table features a column with an example of each consonant in an English word. The letter representing the consonant is highlighted in each case (where possible, it is the first sound of the word).*

Examining Table 2.4, the classic claims about the phonetic properties associated with roundness vs. spikiness are borne out. Obstruents (plosives, affricates, and fricatives) tend to be spiky, whereas sonorants (approximants and nasals) tend to be round. Indeed the ranking closely resembles the sonority hierarchy (i.e. the ranking of syllables by amplitude)<sup>11</sup>, with higher sonority (i.e. amplitude) predicting greater roundness. The major exception to this generalization is that voiced fricatives ([z, v, ʒ]) tend to be spikier than their unvoiced counterparts, and in some cases more spiky than many plosives. There is no readily apparent overall correlation between place of articulation and roundness/spikiness, though it is worth noting that in each

<sup>11</sup> Sonority corresponds to the amount of energy involved in producing a speech sound, i.e. its amplitude, compared to other sounds of equal length, stress, and pitch – Ladefoged, 2001 – and is important in phonological phenomena such as syllable structure. Roughly speaking, the sonority hierarchy, ascending, goes: unvoiced plosive < voiced plosive < unvoiced fricative < voiced fricative < nasal < liquid (e.g. [l, ɹ] – a type of approximant) < semivowel (e.g. [j, w] – a type of approximant) ≤ high vowel < low vowel.

case, the labial consonants ([p, b, m, w]) are the roundest example of their manner of articulation.

Spikiness Ranking	Vowel	Mean Rating	Normalised Rating	Height	Backness	Round ing	Extra Info	Example Word
1	i	4.51	0.49	High	Front			high
2	ɪ	4.45	0.44	High	Front		Lax	pit
3	ʌ	4.41	0.39	Low-mid	Back-Central		Lax	cut
4	ɛ	4.41	0.39	Low-mid	Front		Lax	pet
5	ə	4.38	0.36	Mid	Central		Lax	woman
6	æ	4.29	0.28	Low	Front		Lax	man
7	ɑ	4.29	0.27	Low	Back			car
8	ʊ	4.25	0.23	High	Back		Lax	wood
9	e	4.25	0.23	High-mid	Front			mesa
10	aɪ	4.19	0.18	Low-High	Front		Diphthong	pie
11	ɪə	4.17	0.15	High	Front		Diphthong	here
12	ɒ	4.09	0.07	Low	Back	Round ed	Lax	hot
13	ɜ	4.08	0.06	Low-mid	Central			fur
14	eɪ	4.00	-0.01	High	Front		Diphthong	hay
15	aʊ	3.99	-0.03	Low-High	Back	Round ed	Diphthong	cow
16	ɔ	3.98	-0.04	Low-mid	Back	Round ed		core
17	eə	3.97	-0.05	High	Front		Diphthong	hair
18	ʊə	3.68	-0.34	High	Back		Diphthong	dour
19	ɔɪ	3.67	-0.35	Low-High	Back-Front	Round ed	Diphthong	coy
20	ɑ:	3.66	-0.35	Low	Back		Long	car
21	əʊ	3.65	-0.36	High	Back	Round ed	Diphthong	slow
22	u	3.52	-0.50	High	Back	Round ed		two
23	ɪʊ	3.29	-0.73	High	Front-Back	Round ed	Diphthong	cue
24	ɔ:	3.22	-0.79	Low-mid	Back	Round ed	Long	core

*Table 2.5: Mean syllable scores by vowel. Rows representing syllables that are phonotactically illicit at the start of a word are highlighted in red. The table is arranged in the same way as Table 2.3. Clearly diphthongs present a challenge for a featural coding system, given that they constitute a path through the vowel space rather than a point within it. However, I have represented diphthongs as concatenated monophthongs (as is standard in phonetic transcription), and assigned each diphthong each feature that either of its constituents possesses (in terms of height, backness, and rounding).*

Table 2.5 also supports widespread assumptions about the relationship between phonetics and sound-shape iconicity. Spiky vowels tend to be front or central, whereas round vowels tend to be back vowels or diphthongs with a back component. Round vowels also tend to be phonetically rounded (i.e. labialised), though as this feature is confounded with backness in English, it is difficult to know to what extent each property is contributing to sound-shape iconicity.

### **Statistical Analyses**

Ideally, we might like to have a dataset that covers the entire phonetic space, build an omnibus model for that dataset featuring all distinctive features for our syllables' consonants and vowels, and the interactions between them, and simply see which features and interactions are associated with sound shape iconicity. Unfortunately there are three problems with this approach. Firstly, with an English-specific set of 24 consonants and 24 vowels we do not have enough segments of each type to explore every phonetically possible combination of features, and even our set of over five hundred syllables is not sufficient to explore the whole space of potential interactions between consonant place, consonant manner, consonant voicing, vowel height, vowel backness, and vowel rounding. Secondly, phonetic features are chosen for

phonetic/phonological naturalness, not for optimality as a statistical contrast code. Features are highly intercorrelated, which means choosing between statistically sound predictors and phonetically meaningful ones. Finally, the English phoneme space doesn't come counterbalanced: place and manner are confounded, making it very hard to disentangle their contributions and interactions (e.g. all English dental consonants are also fricatives).

Therefore I will present several more specific analyses, aimed at confirming traditional generalisations about the phonetic correlates of sound-shape iconicity, testing the predictions of specific proposals about the mechanisms of the phenomenon made by other scholars, and testing whether there is any interaction between different kinds of phonetic properties in determining a syllable's iconicity. After each analysis, a version of Table 2.1 or 2.2 will be shown to summarise whether the generalisation (Table 2.1) or prediction (Table 2.2) has been confirmed or not.

**Phonetic correlates of sound-shape iconicity** were confirmed as follows. A t-test showed that as predicted on the basis of previous literature, syllables containing approximants, the most sonorant of consonants ( $M = 3.37$ , 95% CI [3.27, 3.47],  $SD = 0.48$ ), received rounder syllable scores than syllables containing other types of consonant ( $M = 4.14$ , 95% CI [4.09, 4.20],  $SD = 0.63$ ),  $t(159.6) = 13.36$ ,  $p < .001$ , difference = 0.77 (95% CI [0.66, 0.89]), Cohen's  $d = 1.26$ . This result is summarised in Table 2.6.

Round Sound	Source	Spiky Sound	Source
Sonorants 	Ahlner & Zlatev, 2010; Monaghan et al., 2012; Nielsen & Rendall, 2013	Voiceless Obstruents	Ahlner & Zlatev, 2010; D'Onofrio, 2014
Voiced Plosives	D'Onofrio, 2014; Ramachandran & Hubbard, 2001	(Voiceless) Plosives	Aveyard, 2012; Monaghan et al., 2012; Nielsen & Rendall, 2011, 2013; Ramachandran and Hubbard, 2001
Labials	D'Onofrio, 2014		
Back Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2013	Front Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2011
Rounded Vowels	Nielsen & Rendall, 2013	Unrounded Vowels	Nielsen & Rendall, 2013
Low Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012	High Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012

*Table 2.6: The sonorants-are-round generalisation is confirmed*

A second t-test confirmed that syllables containing plosives ( $M = 4.39$ , 95% CI [4.28, 4.50],  $SD = 0.66$ ) received spikier syllable scores than syllables containing other consonant types ( $M = 3.89$ , 95% CI [3.83, 3.95],  $SD = 0.63$ )  $t(306.1) = 7.11$ ,  $p < .001$ , difference = 0.50 (95% CI [0.36, 0.63]), Cohen's  $d = 0.69$ , again, as predicted on the basis of past literature. This is summarised in Table 2.7.

Round Sound	Source	Spiky Sound	Source
Sonorants	Ahlner & Zlatev, 2010; Monaghan et al., 2012; Nielsen & Rendall, 2013	Voiceless Obstruents	Ahlner & Zlatev, 2010; D'Onofrio, 2014
Voiced Plosives	D'Onofrio, 2014; Ramachandran & Hubbard, 2001	(Voiceless) Plosives 	Aveyard, 2012; Monaghan et al., 2012; Nielsen & Rendall, 2011, 2013; Ramachandran and Hubbard, 2001
Labials	D'Onofrio, 2014		
Back Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2013	Front Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2011
Rounded Vowels	Nielsen & Rendall, 2013	Unrounded Vowels	Nielsen & Rendall, 2013
Low Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012	High Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012

Table 2.7: The plosives-are-spiky generalisation is confirmed

A t-test confirmed that syllables containing back vowels received rounder syllable scores ( $M = 3.77$ , 95% CI [3.69, 3.86],  $SD = 0.68$ ) than those containing front or central vowels ( $M = 4.26$ , 95% CI [4.19, 4.33],  $SD = 0.57$ ),  $t(544.1) = 9.13$ ,  $p < .001$ , difference = 0.48 (95% CI [0.38, 0.59]), Cohen's  $d = 0.77$ . Any diphthong with a back component was counted as a back vowel. Again this reflected results presented in past literature. This result is summarised in Table 2.8.

Round Sound	Source	Spiky Sound	Source
Sonorants	Ahlner & Zlatev, 2010; Monaghan et al., 2012; Nielsen & Rendall, 2013	Voiceless Obstruents	Ahlner & Zlatev, 2010; D'Onofrio, 2014
Voiced Plosives	D'Onofrio, 2014; Ramachandran & Hubbard, 2001	(Voiceless) Plosives	Aveyard, 2012; Monaghan et al., 2012; Nielsen & Rendall, 2011, 2013; Ramachandran and Hubbard, 2001
Labials	D'Onofrio, 2014		
Back Vowels 	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2013	Front Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2011
Rounded Vowels	Nielsen & Rendall, 2013	Unrounded Vowels	Nielsen & Rendall, 2013
Low Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012	High Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012

Table 2.8: The back vowels-are-round generalisation is confirmed

In order to explore whether there was any evidence for an effect of rounding over and above backness, I constructed two linear mixed effects models for the data using the R package lme4 (Bates, Maechler, Bolker, & Walker, 2015), one featuring only backness as a predictor, one featuring both backness and rounding, and both featuring random intercepts for consonant identity, and random slopes for all of their predictors by consonant identity. Both backness and rounding were coded +0.5 vs. -0.5. The latter model suggested that both backness ( $\beta = -0.282$ , 95% CI [-0.405, -0.158],  $t = 4.21$ ) and rounding ( $\beta = -0.245$ , 95% CI [-0.371, -0.119],  $t = 3.68$ )<sup>12</sup> contribute to iconic roundness. I then tested whether a model with the extra predictor of rounding was a significant improvement on the more restricted model using R's *anova* function, which compares nested models using the log likelihood of the

<sup>12</sup> Confidence intervals were computed using R's *confint* function.

dataset given each model. The comparison suggests that the model featuring roundness is a significant improvement on the model featuring backness only ( $\text{Chi}^2 = 15.67, p = .004$ ). This result is summarised in Table 2.9.

Round Sound	Source	Spiky Sound	Source
Sonorants	Ahlner & Zlatev, 2010; Monaghan et al., 2012; Nielsen & Rendall, 2013	Voiceless Obstruents	Ahlner & Zlatev, 2010; D'Onofrio, 2014
Voiced Plosives	D'Onofrio, 2014; Ramachandran & Hubbard, 2001	(Voiceless) Plosives	Aveyard, 2012; Monaghan et al., 2012; Nielsen & Rendall, 2011, 2013; Ramachandran and Hubbard, 2001
Labials	D'Onofrio, 2014		
Back Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2013	Front Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2011
Rounded Vowels ✓	Nielsen & Rendall, 2013	Unrounded Vowels ✓	Nielsen & Rendall, 2013
Low Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012	High Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012

*Table 2.9: The rounded vowels-are-round generalisation is confirmed (over and above the effect of backness)*

As stated in the introduction, some pieces of literature seem to imply that high vowels are spiky and low vowels are round (Monaghan, Mattock, & Walker, 2012, in their choice of vowels in Experiment 2; Ahlner & Zlatev, 2010, p. 324), though this is not nearly so widespread as the claim that vowel backness affects sound-shape iconicity. Intuitively the relationship between vowel height and shape might make sense, given that low vowels imply a bigger opening in the oral cavity, and larger size seems to be associated with roundness. However, we find evidence that the opposite relationship actually holds, if anything. Excluding diphthongs with both a

high and low component, syllables with high vowels (as coded according to Hall, 2007, who uses two height features: high and low) received lower (i.e. rounder) syllable scores ( $M = 3.92$ , 95% CI [3.83, 4.01],  $SD = 0.70$ ) than syllables with low vowels ( $M = 4.15$ , 95% CI [4.03, 4.27],  $SD = 0.66$ ):  $t(249.1) = 2.99$ ,  $p = .003$ , difference = 0.23 (95% CI [0.08, 0.38]) Cohen's  $d = 0.33$ . It should not however be concluded that all high vowels are round, as backness is also influential (think of [i], [ɪ]). Is it possible that the height result is simply the result of more high than low vowels being back vowels?

Follow up models confirm that the influence of height is not simply attributable to a confound with backness. I constructed two linear mixed effects models for the data using lme4, one featuring only backness as a predictor, one featuring both backness and the two height features as predictors: high, and low. Both models featured random intercepts for consonant identity, and random slopes for all of their predictors by consonant identity. Both backness and the height variables were coded +0.5 vs. -0.5. A high vowel would be +high, -low, a low vowel + low, -high, and a mid vowel – low –high. The model with height predictors suggested that backness ( $\beta = -0.543$ , 95% CI [-0.618, -0.467],  $t = -14.10$ ) contributes to iconic roundness, whereas lowness ( $\beta = 0.263$ , 95% CI [0.165, 0.362],  $t = 5.27$ ) contributes (less strongly) to iconic spikiness (there was no overall difference in spikiness between mid and high vowels). I then tested whether a model with the extra predictors for height was a significant improvement on the more restricted model using R's *anova* function. The comparison suggests that it is ( $\text{Chi}^2 = 33.23$ ,  $p < .001$ ). Note that this cannot be attributed to the fact the high front vowels are more fronted than low front vowels, as this would predict an effect in the opposite direction (i.e. spikier high vowels). This failure to confirm is summarised in Table 2.10.

Round Sound	Source	Spiky Sound	Source
Sonorants	Ahlner & Zlatev, 2010; Monaghan et al., 2012; Nielsen & Rendall, 2013	Voiceless Obstruents	Ahlner & Zlatev, 2010; D'Onofrio, 2014
Voiced Plosives	D'Onofrio, 2014; Ramachandran & Hubbard, 2001	(Voiceless) Plosives	Aveyard, 2012; Monaghan et al., 2012; Nielsen & Rendall, 2011, 2013; Ramachandran and Hubbard, 2001
Labials	D'Onofrio, 2014		
Back Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2013	Front Vowels	Ahlner & Zlatev, 2010; D'Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2011
Rounded Vowels	Nielsen & Rendall, 2013	Unrounded Vowels	Nielsen & Rendall, 2013
Low Vowels 	Ahlner & Zlatev, 2010; Monaghan et al., 2012	High Vowels 	Ahlner & Zlatev, 2010; Monaghan et al., 2012

*Table 2.10: The low vowels-are-round generalisation is falsified – if anything the reverse is true*

Regarding specific predictions about the effect of voicing of obstruents (Ahlner & Zlatev, 2010; D'Onofrio, 2014; Ramachandran & Hubbard, 2001), these were best tested as part of the models I report under the interaction heading below.

**Previous claims** by Fort, Martin, & Peperkamp (2015) and Cuskley, Simner, & Kirby (2015) were followed up. Fort et al., working with Francophone participants, found that consonants are a more powerful influence than vowels in determining which of a rounded and a spiky shape participants choose to pair with a nonword. To see whether I found a similar pattern, I constructed two very simple models: one where every syllable's syllable score was modelled as the mean for its consonant, and one

where it was modelled as the mean for its vowel<sup>13</sup>. As predicted, the consonant model explained appreciably more of the variance than the vowel model (52% vs. 28%), suggesting that as with Fort et al.'s study, consonants were more influential in determining sound-shape iconicity than vowels.

Cuskley et al. argue that shape of the corresponding letter is an important determinant of the iconic properties of phonemes. They found that in terms of spikiness ratings, [k, t] > [z, v] > [s, f] > [d, g]. Our findings largely coincide with this, except that our participants deemed [d] spikier than [s, f], and [g] spikier than [s, f, v]. Clearly some of these relations are not what is predicted by letter shape, and seem more suited to an explanation in terms of a phonetic property like sonority<sup>14</sup>.

Turning to phonemes that are consistently represented with identical or near identical letterforms, which Cuskley et al. would predict to be equally round/spiky, we find that there is a significant difference between syllables containing [b] ( $M = 3.79$ , 95% CI [3.60, 3.97],  $SD = 0.44$ ) and syllables containing [p] ( $M = 4.18$ , 95% CI [3.93, 4.43],  $SD = 0.60$ ),  $t(42.2) = 2.64$ ,  $p = .01$ , difference = 0.40 (95% CI [0.09, 0.70]), Cohen's  $d = 0.76$ . However, though [ð] and [θ] are both represented by the orthographic double *th*, syllables containing the voiced dental fricative [ð] ( $M = 4.07$ , 95% CI [3.88, 4.27],  $SD = 0.43$ ) are not significantly spikier than those containing the voiceless [θ] ( $M = 4.02$ , 95% CI [3.84, 4.21],  $SD = 0.44$ ),  $t(42.4) = 0.4$ ,  $p = .701$ <sup>15</sup>. I followed up with a Bayesian t-test using the R package *BayesFactor* (Morey & Rouder, 2015), comparing an  $H_0$  that there is no difference between the groups to an

---

<sup>13</sup> Note that while this comparison would be too simplistic if the data were structured differently, it is perfectly appropriate here given that there are an equal number of consonants and vowels, and no correlation between their assignment to syllables.

<sup>14</sup> Note though that as observed above, a sonority-based account has difficulty explaining why voiced fricatives are spikier than their voiceless counterparts.

<sup>15</sup> N.B. the same result holds if the t-test is paired by vowel identity:  $t(20) = 0.27$ ,  $p = .792$ .

$H_1$  corresponding to an uninformative Cauchy prior on effect size<sup>16</sup>. This test yielded a Bayes factor of 3.19 in favour of the null, providing the null with moderate support (Jeffreys, 1961; Kass & Raftery, 1995; see Chapter 3 for more on the interpretation and value of Bayesian statistics). This means that the lack of a significant difference reflects evidence in favour of the null rather than limited power.

Table 2.11 summarises the fact that there is mixed evidence for the orthography account of sound-shape iconicity.

Theory	Round Predictions	Spiky Predictions
Cross-modal analogy	Smooth sounds – approximants?	Sharp sounds - plosives?
Cross-wiring	?	?
Lip shape	Labial/rounded: [p, b, m, w, o, u]	Everything else?
Orthography  	[w, b, p, s, d, m, o, u]	[k, z, t, v, i]
Environmental statistics	Low pitches?	High pitches?

Table 2.11: Some predictions of the orthography account are borne out, but not all

**Interactions between consonant place, consonant manner, and vowel** were investigated as follows.

<sup>16</sup> Centred on zero with a scale of  $0.5^{0.5}$ , and with effect size standardized such that 1 = the difference between the group means divided by the pooled standard deviation.

As discussed above, an omnibus model of the dataset containing huge numbers of predictors and interactions is not feasible. However, I will make use of some regularities in the phoneme inventory of English to make it tractable to look at interactions between place, manner, and vowel for some subclasses of English consonant. Firstly, English has the same trio of stops at three different places of articulation:

	<b>Bilabial</b>	<b>Alveolar</b>	<b>Velar</b>
<b>Voiceless Plosive</b>	[p]	[t]	[k]
<b>Voiced Plosive</b>	[b]	[d]	[g]
<b>Nasal Stop</b>	[m]	[n]	[ŋ]

*Table 2.12: Inventory of English stops*

The fact that this set is complete means that place and manner are not confounded within it, obviating the problems we have with the consonant set as a whole and making it possible to analyse their interaction. Similarly English has the same pair of fricatives, manner-wise, at four separate places of articulation:

	<b>Labiodental</b>	<b>Dental</b>	<b>Alveolar</b>	<b>Postalveolar</b>
<b>Voiceless Fricative</b>	[f]	[θ]	[s]	[ʃ]
<b>Voiced Fricative</b>	[v]	[ð]	[z]	[ʒ]

*Table 2.13: (Partial) inventory of English fricatives*

These too can be analysed for interactions between place and manner.

The stop set was orthogonally contrast coded (for place, and separately for manner) as follows:

Phoneme	PLACE		MANNER		VOWEL
	Bilabial	Alveolar	Nasal	Voiced	Back
[p]	0.67	0	-0.33	-0.5	0.5/-0.5
[b]	0.67	0	-0.33	0.5	0.5/-0.5
[m]	0.67	0	0.67	0	0.5/-0.5
[t]	-0.33	0.5	-0.33	-0.5	0.5/-0.5
[d]	-0.33	0.5	-0.33	0.5	0.5/-0.5
[n]	-0.33	0.5	0.67	0	0.5/-0.5
[k]	-0.33	-0.5	-0.33	-0.5	0.5/-0.5
[g]	-0.33	-0.5	-0.33	0.5	0.5/-0.5
[ŋ]	-0.33	-0.5	0.67	0	0.5/-0.5

*Table 2.14: Contrast codes for the stop model. The “Vowel” column reflects the fact that individual syllables were coded as +0.5 if they featured a back vowel, -0.5 if they did not.*

The model, using lme4, comprised all the above predictors, plus interactions between predictors from different categories (e.g. place and vowel). I also included random intercepts, and random slopes for place and manner predictors by vowel identity (see Appendix 2.6 for full model specification).

Starting with main effects: for place, there were reliable effects of the bilabial code ( $\beta = -0.468$ , 95% CI [-0.554, -0.382],  $t = -10.62$ ), indicating that syllables containing bilabial stops tend to be rounder than others, and of the alveolar code ( $\beta = -0.202$ , 95% CI [-0.303, -0.102],  $t = -3.95$ ), indicating that alveolar stops tend to be rounder than velars. For manner, there were reliable effects of both the nasal code ( $\beta = -0.814$ , 95% CI [-0.917, -0.712],  $t = -15.60$ ), indicating that nasals are rounder than plosives, and of the voicing code ( $\beta = -0.582$ , 95% CI [-0.703, -0.461],  $t = -9.40$ ), indicating that voiced plosives tend to be rounder than their voiceless counterparts. There was also a main effect of backness ( $\beta = -0.509$ , 95% CI [-0.751, -0.267],  $t = -$

4.13), indicating that syllables containing back vowels tended to be rounder than those that did not.

The fact the bilabial stops are especially iconically round, and the fact that rounded vowels are also especially iconically round (as established earlier) bear out the predictions of the lipshape theory. However there are other salient facts about the dataset that the lipshape theory cannot explain: the effect voicing has on the roundness of bilabial stops, the roundness of non-labial consonants like [j], and the effect of backness on vowel roundness independently of lip rounding. Thus the lipshape account only receives partial support from the data, a state of affairs summarized in Table 2.15.

Theory	Round Predictions	Spiky Predictions
Cross-modal analogy	Smooth sounds – approximants?	Sharp sounds - plosives?
Cross-wiring	?	?
Lip shape  	Labial/rounded: [p, b, m, w, o, u]	Everything else?
Orthography	[w, b, p, s, d, m, o, u]	[k, z, t, v, i]
Environmental statistics	Low pitches?	High pitches?

*Table 2.15: The basic prediction of the lipshape account that there is something particularly round about labials is confirmed, but the account fails to predict other patterns in the effect*

Moving to interactions, there was a reliable interaction between the bilabial code and the nasal code ( $\beta = 0.388$ , 95% CI [0.218, 0.558],  $t = 4.47$ ), indicating that the difference in roundness between the nasal and the plosives was less pronounced for bilabial stops than for other places of articulation. There was also a reliable interaction between the bilabial code and the voicing code ( $\beta = 0.277$ , 95% CI [0.081, 0.473],  $t = 2.77$ ), indicating – similarly – that the difference in roundness between [b] and [p] was less pronounced than the distinction between the voiced and voiceless plosive at other places of articulation. Furthermore, there was a reliable interaction between the alveolar code and the nasal code ( $\beta = 0.427$ , 95% CI [0.23, 0.625],  $t = 4.24$ ), meaning that distinction between nasals and plosive was smaller for the alveolar than for the velar stops. Finally, there was an interaction between the bilabial code and the backness code ( $\beta = -0.188$ , 95% CI [-0.36, -0.015],  $t = -2.13$ ), showing that, on average, pairing the stop with a back vowel added more roundness for bilabial stops than for alveolars or velars.

The fricative set was contrast coded as follows (again, orthogonally within each dimension):

Phoneme	PLACE			VOICING	VOWEL
	Labiodental	Dental	Alveolar	Voiced	Back
[f]	0.75	0	0	-0.5	0.5/-0.5
[v]	0.75	0	0	0.5	0.5/-0.5
[θ]	-0.25	0.67	0	-0.5	0.5/-0.5
[ð]	-0.25	0.67	0	0.5	0.5/-0.5
[s]	-0.25	-0.33	0.5	-0.5	0.5/-0.5
[z]	-0.25	-0.33	0.5	0.5	0.5/-0.5
[ʃ]	-0.25	-0.33	-0.5	-0.5	0.5/-0.5
[ʒ]	-0.25	-0.33	-0.5	0.5	0.5/-0.5

*Table 2.16: Contrast codes for the fricative model. The “Vowel” column reflects the fact that individual syllables were coded as +0.5 if they featured a back vowel, -0.5 if they did not.*

The model was constructed in the same way as the last, with interactions between predictors from different categories included, and random effects by vowel identity (with slopes for main effects but not interactions). See Appendix 2.7 for full model specification.

Reporting main effects first: There was a reliable effect of the dental predictor ( $\beta = -0.157$ , 95% CI [-0.296, -0.017],  $t = -2.2$ ), indicating that syllables with dental fricatives tended to be rounder than those with alveolar or postalveolar fricatives. There was also a main effect of the alveolar predictor ( $\beta = 0.243$ , 95% CI [0.098, 0.388],  $t = 3.29$ ), indicating that [t] and [d] syllables tend to be spikier than [ʃ] and [ʒ]. There was a main effect of the voice predictor ( $\beta = 0.267$ , 95% CI [0.157, 0.376],  $t = 4.77$ ), indicating that voiced fricatives tended to be spikier than voiceless equivalents. There was also a main effect of backness ( $\beta = -0.429$ , 95% CI [-0.637, -0.222],  $t = -4.05$ ), meaning that syllables with back vowels tended to be rounder than those without.

Turning to interactions: the interaction between the dental predictor and the voicing predictor is significant ( $\beta = -0.369$ , 95% CI [-0.579, -0.158],  $t = -3.43$ ), indicating that voicing contributes less spikiness with dental fricatives than with alveolar and postalveolar fricatives (in keeping with the earlier analysis of dental fricatives). Finally, there was a reliable interaction between the alveolar predictor and the voicing predictor ( $\beta = 0.508$ , 95% CI [0.269, 0.747],  $t = 4.17$ ), demonstrating that voicing adds more spikiness for alveolar fricatives than for postalveolar fricatives

(this may reflect the fact that spectrograms typically show weaker voicing bars for [ʒ] than for [z]: Ladefoged, 2001, p. 183).

Note that as predicted by much of the earlier literature (Ahlner & Zlatev, 2010; D’Onofrio, 2014; Ramachandran & Hubbard, 2001), voiced plosives are rounder than voiceless plosives, but that this pattern is however *reversed* for fricatives. This more nuanced version of the previously reported pattern is summarized in Table 2.17.

Round Sound	Source	Spiky Sound	Source
Sonorants	Ahlner & Zlatev, 2010; Monaghan et al., 2012; Nielsen & Rendall, 2013	Voiceless Obstruents	Ahlner & Zlatev, 2010; D’Onofrio, 2014
Voiced Plosives	D’Onofrio, 2014; Ramachandran & Hubbard, 2001	(Voiceless) Plosives	Aveyard, 2012; Monaghan et al., 2012; Nielsen & Rendall, 2011, 2013; Ramachandran and Hubbard, 2001
Labials	D’Onofrio, 2014		
Back Vowels	Ahlner & Zlatev, 2010; D’Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2013	Front Vowels	Ahlner & Zlatev, 2010; D’Onofrio, 2014; Monaghan et al., 2012; Nielsen & Rendall, 2011
Rounded Vowels	Nielsen & Rendall, 2013	Unrounded Vowels	Nielsen & Rendall, 2013
Low Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012	High Vowels	Ahlner & Zlatev, 2010; Monaghan et al., 2012



*Table 2.17: The generalisation that voiceless obstruents are particularly spiky is only true for plosives – the reverse is true for fricatives*

Thus to recap, our models showed most of the expected effects of manner, voicing (though see the last paragraph), and vowel, and more interestingly, effects of place

of articulation, and some interactions between different types of phonetic properties (e.g. place and manner).

## Discussion

I had monolingual Anglophone participants rate a near exhaustive set of the CV syllables that can be generated using the English phonemic inventory. A mean rating (aka syllable score) was calculated for each syllable, which I used to explore the data in a variety of ways. This both sets the scene for the next two chapters, and represents a novel contribution to the sound-shape iconicity literature.

### **Basic Outline of Sound-Shape Iconicity**

First I ranked both consonants and vowels by roundness vs. spikiness. My results provide newly comprehensive evidence to support most longstanding generalizations about the phonetic basis of sound-shape iconicity. High sonority consonants like approximants and nasals tend to suggest roundness, whereas certain kinds of low sonority consonants like voiceless stops and voiced fricatives suggest spikiness. Bilabial consonants appear to be particularly round (the roundest consonant of each manner is a bilabial, where there is a bilabial version of that manner). Back vowels tend to be round, whereas front and (the small number of) central vowels tend to be spiky. Rounded vowels (i.e. those produced with rounded lips) also tend to receive round syllable scores, though rounding is confounded with backness in English.

Following up with inferential statistics, I found that as predicted on the basis of previous work, approximants are rounder than other consonants, stops are spikier than other consonants, and back vowels are rounder than other vowels. Analysis of the later stops model also showed that voiced stops are rounded than voiceless stops, again as predicted. A model containing predictors for both backness and rounding proved to be a significant improvement on a model containing the predictor of backness alone, suggesting that in spite of their being confounded, rounding tends to contribute to round iconicity over and above backness. Vowel height was a (relatively weak) predictor of sound-shape iconicity, but in the *opposite* direction to that suggested by Monaghan, Mattock, and Walker (2012), and Ahlner and Zlatev (2010): low vowels are *less* round than others. Moreover the later fricatives model shows that voicing makes fricative *more* rather than less spiky, thus the effect of voicing on obstruents depends on whether the obstruent is a fricative or a stop, contrary to some previous generalisations about voicing making all consonants or all obstruents rounder (Ahlner & Zlatev, 2010; D’Onofrio, 2014).

English’s preference for rounding back vowels is not a coincidence. Cross-linguistically there are many languages (like French) that have rounded front and back vowels, and many languages (like English) that have rounded back vowels without rounded front vowels, but exceedingly few (perhaps none) that have rounded front vowels without rounded back vowels (Kaye, 1989). This reflects the fact that backness and rounding both primarily affect the same aspect of acoustic phonetics by lowering the second formant. Having rounded back vowels effectively expands the phonetic space backwards, making back vowels more perceptually discriminable (a factor that seems to systematically shape vowel systems – Liljencrants & Linblom, 1972 – perhaps via evolutionary pressures on language change). Thus it may be that

rather than being alternative sources of roundness, rounding and backness both contribute to an acoustic property which is the direct cause of the sense of roundness for vowels (this is not to presuppose an answer to the question of *why* that phonetic property is associated with roundness – quite possibly the answer to this will involve its being diagnostic of backness, rounding, or both).

### **Specific Claims Assessed**

I then performed analyses aimed at testing interesting claims from the recent literature about the basis of sound-shape iconicity. In keeping with Fort, Martin, & Peperkamp's (2015) finding that consonants influence the sound-shape iconicity properties of a word more than vowels, I found that a model only taking into account consonants captured more of the variance in the dataset than one only taking into account vowels. All my syllables were of the structure CV, meaning that in my dataset consonant vs. vowel is confounded with primacy of phoneme in the utterance. Fort et al. did control for this in their earlier study and continued to find that consonants made the more important contribution. However, it would be interesting to investigate how much this might reflect speech perception devoting more resources to consonants than vowels because the former carry more information about lexical identity (Owren & Cardillo, 2006; though note that this is controversial: Kewley-Port, Burkle, & Lee, 2007).

I also followed up on Cuskley, Simner, and Kirby's (2015) finding that the shape of the letter used to represent a consonant is a strong predictor of that consonant's sound-shape iconicity, even in a spoken context, a result that they use to argue that to a large extent the kiki-bouba effect is based on orthography. I found some

patterns in the data that were consistent with this claim, but others that were less so. For instance, [p] was significantly spikier than [b], in spite of them being (reflected and rotated) versions of the same letterform. The [b, p] distinction is clearly hard to explain on the assumption that a phoneme's iconicity is purely based on its own individual orthographic representation, though it could be explained if phonemic iconicity is influenced by the letterforms of other phonemes in the same natural class (e.g. while *b* is no rounder than *p*, *g* is rounder than *k*, a distinction that may end up influencing [b, p] as well as [g, k], as both are pairs of voiced vs. voiceless stops at a given place of articulation). Likewise, the ranking of consonants I obtained is not entirely consistent with Cuskley et al.'s findings, with [g] receiving spikier syllable scores than [f] and [v].

On the other hand, I found evidence that there is no difference between the syllable scores of [ð] and [θ], the only two consonants that almost invariably receive the same orthographic value (i.e. *th*). However, the linguist who recorded my stimuli reported that it was difficult to keep [ð] and [θ] distinct, perhaps because they are not very perceptually discriminable (indeed anecdotally lay English speakers generally have no sense that *th* covers two separate speech sounds), which may explain why syllable scores differ less than those between other pairs of voiced and voiceless fricatives.

### **Interactions between Place and Manner and Implications for the Mechanism of Sound-Shape Iconicity**

Finally, I exploited regularities in the consonant inventory of English to look at groups of syllables that show the natural equivalent of a counterbalanced design, crossing

place and manner features. This allowed me to explore the extent to which different aspects of my syllables (specifically place and manner, place and vowel, and manner and vowel) interacted in determining the iconic properties of the syllable. I found some evidence for interactions between place and manner, and place and vowel type. In particular, the model for stops (both plosives and nasals) shows that bilabial stops are not only rounder than alveolars and velars overall, but that there is less variation between the iconic properties of different manners in the bilabial position than in other positions. This, along with the observation that consonants of a given manner articulated bilabially tend to be rounder than consonants of the same manner articulated elsewhere, raises the question of what might be special about the lips.

There are several possibilities. One is based on letterform: *p* is at least as round as *t* or *k*; *b* is as round as *d*, though perhaps not *g*; *m* is as round as *n* or *ng*; and *w* is as round as *y*, *r*, or *l*. Thus overall the letterforms associated with bilabial consonants are rounder than their neighbours by manner, though this does not explain why our participants found [b] so much rounder than [g].

Another story would be that the consonants' bilabiality *per se* is responsible for their iconic roundness. If speakers of a language come to associate phonemes with visual, motoric, or proprioceptive representations of the articulatory gestures used to produce them, then it is possible that bilabial sounds are preferentially mapped onto round objects via intermediary representations of roundish pairs of lips (even closed lips are reasonably rounded, and [w] involves literal rounding of the lips). This opens up the interesting possibility that at least some parts of sound-shape iconicity are based on unimodal mappings directly from object to lip shape, and that this aspect of the effect is learned from linguistic input without being language specific. As the

articulatory system is universal, then speakers of any language that uses these bilabial phonemes will develop the same kinds of iconic associations. This though fails to explain why [j], which does not involve the lips, appears to be the second roundest consonant.

Yet another possibility is that the explanation lies in the hallmarks of these bilabials' acoustic phonetics. A number of the bilabial consonants acoustically resemble the rounded back vowels: the most iconically round of vowels. Typically, speech takes the form of alternation between consonants and vowels, and acoustically speaking consonants can be thought of as ways of inflecting the end of the last vowel and/or the beginning of the next (Ladefoged, 2001). Firstly, the consonant [w] is essentially a consonantal version of the high back rounded vowel. A CV syllable beginning with [w] essentially takes the form of a glide to the vowel from a very short version of [u]. Secondly, the place of articulation of a stop is mainly distinguishable on the basis of the second and third formant transitions to and from surrounding vowels (Ladefoged, p. 179-180). Both [n] and [b] are characterised by formant transitions where the locus (i.e. the apparent place of origin within the consonant) of the second formant is comparatively low. As already discussed, a low second formant is the mark of back rounded vowels. Finally, [p], as well as sharing the second formant transition typical of [b] and [m] (albeit in weaker form, as voicing is absent), is also quieter than [t] and [k], featuring much less of the high frequency energy associated with aspiration (Ladefoged, p. 181). Thus in as much as the sound of aspiration carries the spikiness of sounds like [t] and [k] (either because of a general mapping between high frequency sounds and small, quick, spiky things, or because it is diagnostic of unvoiced stops, which have spiky associations for other reasons), we would predict that [p] is less spiky than [k] or [t], which is what we find.

These observations hold out the possibility of subsuming at least some aspects of both consonant and vowel iconicity under a shared phonetic account. Basically, the observation is that consonants whose acoustic phonetics resemble back vowels (e.g. [w], [m], [b], [r], and [l], the latter two of which closely resemble [w] viewed spectrographically) tend to be round, whereas consonants whose formant transition acoustics make them resemble front vowels (e.g. [k] and [t]) are spiky. The question would then be why back vowels should suggest roundness and front vowels spikiness. This might turn out to be explicable in terms of the amount of energy at different frequencies in these vowel, and a non-linguistic tendency to associate high frequencies with smaller sharper things, which would in turn be based on statistical correlations present in sensory input (Marks, 1987; see Spence, 2011, for a review of such cross-modal correspondences).

Additionally (this possibility is not necessarily mutually exclusive), the association may be based on what Ohala (1994) describes as the *frequency code* – that is the propensity of certain features of animal vocalizations to carry information about the size of the animal, a propensity that animals manipulate to appear bigger or smaller in the right circumstances. In particular, formants are a good proxy for the length of the vocal tract, which in turn is a good proxy for the size of the animal. Low formants signal a big beast: this is why red deer stags lower their larynx when calling (Fitch & Reby, 2001). It is also why, Ohala suggests, humans smile when they wish to appear agreeable, and why other mammal species retract the lips when submissive: this gesture shortens the vocal tract to signal a smaller, less threatening animal. If, for independent reasons (perhaps statistical correlations in the environment), roundness and largeness, and spikiness and smallness, are associated (Perniss, Thompson, & Vigliocco, 2010), then we have a neat account of why phonemes with low formants,

i.e. rounded back vowels and the consonants that resemble them, should suggest roundness, and why their front counterparts should suggest spikiness.

However, this account runs into two immediate problems. The first is why  $F_2$  should be so much more important in the phenomenon than  $F_1$ : both decrease in pitch when the length of the vocal tract is extended (Fitch & Reby, 2001; though note that we found that  $F_2$ , i.e. vowel height, *does* show an effect in the right direction, albeit a much smaller one than backness). Even if this can be explained, we are still left with the puzzle of why a consonant like [j], the consonantal equivalent of [i] i.e. the frontmost vowel in English, should be considered highly round by our participants. This might be rectified by adding sonority as a second correlate of roundness, but even this would fail to explain why voiced fricatives should be considered spikier than the voiceless equivalents they only differ from by the addition of low frequencies.

## **Conclusion**

In truth, it appears that no single principle is sufficient to capture the patterns we see in this dataset. It seems likely that a number of factors are at work in determining the sound-shape iconicity of a phoneme, possibly including associated letterform, associated lip shape, pitch profile, and sonority. There are however particular types of evidence that can help verify whether each of these factors is in play. If letterform is important then we would expect the iconic associations of a phoneme to vary depending on speakers orthographic systems. If any of the phonetics-based accounts are correct, then we would expect non-linguistic sounds with similar acoustic profiles to a phoneme to have the same iconic connotations. And if lip

shape is important then we would expect an asymmetry in the effect, given that it is partly driven by a rounded-round pairing that has no obvious spiky equivalent. This asymmetry may have been masked by the binary forced choice methodologies typically used to investigate sound-shape iconicity so far, but should show in other paradigms. We will return to this last question in Chapters 3 and 4.

One obvious limitation of the study is that it only used literate adult Anglophones. While this is my population for most of the experiments in subsequent chapter, meaning that this study does succeed in laying the groundwork for assumptions about their sensitivity to sound-shape iconicity, it does mean that my findings may not generalise to speakers of other languages, children and infants, and people who cannot read and write. These will all be important populations to investigate further as scholars learn more about the boundaries and basis of the effect.



# Chapter 3: Cross-Situational Learning<sup>17</sup>

## Introduction

This chapter uses cross-situational learning, an artificial language learning paradigm that mirrors natural language input by teaching participants vocabulary incrementally over multiple exposures. The experiments focus on sound-shape iconicity. Our thesis is that if iconicity helps resolve referential ambiguity, then iconically congruent vocabulary should be learned more readily than iconically incongruent vocabulary. Experiment 1 is a near-replication of Monaghan, Mattock, & Walker (2012), that largely reproduces their findings but diverges from them in one interesting respect. Experiments 2-4 modify the same paradigm to investigate whether round and spiky iconicity are the same strength – an important step towards establishing the mechanism of the effect.

## The Paradigm

Cross-situational learning grew out of the vocabulary acquisition literature. A long running topic of research in that field concerns the sources of information young children use to single out word meanings. One view is that they wait until they encounter a word in a context where its meaning is very low in ambiguity, and only then posit a possible meaning for the word by 'fast mapping' (Carey & Bartlett, 1978) with the help of attentional (Smith, 2000), conceptual (Gentner, 1982), linguistic (Gleitman, 1990) and social (Baldwin, 1993; Bloom, 2000, Tomasello, 2000) biases and knowledge. Under this theory a word-learning event might take the form of a

---

<sup>17</sup> Thank you to Zoë Belk and Pamela Perniss for recording stimuli

parent saying “hi baby, can you pass me the duck”, at a moment when a duck is the only salient object in the child’s visual field. Assuming the right knowledge and biases - that the child knows the other words in the sentence; that it knows that “duck” must be a noun; that the child is biased towards interpreting nouns as referring to whole objects (rather than meaning e.g. “undetached duck part”) the input at the moment of exposure to the word unambiguously links the word with its meaning.

An alternative view is that children integrate information across multiple exposures to a word (Siskind, 1996; Vogt & Smith, 2005; Yu & Ballard, 2007). This increases the power of word learning because even a set of individually ambiguous exposures might be unambiguous when taken together. *Prima facie* support for this view includes the fact that even under optimal lab conditions, children as old as 18 months (i.e. well into lexical acquisition) often have difficulty inferring the meaning of a word in a single exposure (Hirsh-Pasek, Golinkoff, & Hollich, 1999; Moore, Angelopoulos, & Bennett, 1999; Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2006). The picture of word-learning on this view is that the child experiences a number of exposures to the word “duck”, each featuring not only a duck but also other objects (and possibly at times no duck at all). Though no single event suffices to teach the child the word, if she is keeping track of the number of times each kind of referent appears with the word (perhaps through repeatedly updated association strengths) then she may end up with the correct referent as the clear winner.

The cross-situational learning paradigm was originally designed by Chen Yu and Linda Smith as a way to test this second account experimentally (Yu & Smith, 2007). The idea is to expose participants to a series of artificial naming events that are ambiguous individually, but informative en masse through co-occurrence

relationships. Learning from this kind of information is feasible in principle (Siskind, 1996; Vogt & Smith, 2005; Yu & Ballard, 2007), but had not been previously been demonstrated in humans. Yu and Smith showed that given trials where two, three, or four pictures of unusual objects were presented along with invented names for each (with 18 names and objects in total), six exposures to each name saw adult participants performing well above chance. Smith and Yu (2008) extend this finding to 12- and 14-month-old infants. Infants were shown training slides featuring pairs of novel pictures accompanied by invented names for the pictures, played sequentially. There were six word-referent pairings in all, and each correct word-referent pairing was presented ten times for each infant. The training phase was followed by testing: over the course of 12 trials each picture was presented along with its name and a distractor (another one of the pictures, this one left unnamed). Both age groups looked for significantly longer at targets than at distractors, indicating that they had used cross-situational statistics to learn about word-referent pairings (a result replicated in Yu & Smith, 2011). Yurovsky, Smith, and Yu (2013) used headcams to capture the point of view of 2-to-2.5-year-old children during a naturalistic free play session with their parents. Yurovsky et al. then took the most ambiguous naming events, replaced the toy's name with an invented word, and used the footage as slides in a cross-situational learning experiment for adults. Adults successfully learned word meanings, establishing the plausibility of cross-situational learning being a powerful vocabulary acquisition tool when applied to the kind of input young children receive outside the lab.

The mechanism underlying cross-situational learning is not yet clear. There are at least two possibilities: a comparatively 'dumb' associational learning mechanism, which keeps track of co-occurrence rates, or a mechanism that tests explicit and

coherent hypotheses. However, as Yu and Smith (2012) demonstrate, each of these broad classes of mechanisms can end up predicting very similar patterns of learning depending on the details of the specific models. Thus resolving this question must await richer data and more constrained models.

Since Yu and Smith's first papers, cross-situational learning has been adapted for adult use (e.g. Monaghan, Mattock, Davies, & Smith, 2015). It has two advantages as an artificial language learning paradigm. First of all it's naturalistic in that it involves no explicit ostensive teaching of vocabulary: the research above shows that it can be used by children (and adults) to learn word meanings, and that it very plausibly *is* used this way during development. Secondly, in the form we use below (following Monaghan, Mattock, & Walker, 2012), each trial constitutes both learning and testing: participants are exposed to a word and to possible referents, and then have to select the referent that matches the word (guessing at first, and learning more and more as they see more trials). The result is that the paradigm is very well suited to gauging the rate of learning over training trials.

### **Previous Application to Sound-Shape Iconicity**

Monaghan, Mattock, and Walker (2012) use the paradigm to explore whether and how sound-shape iconicity boosts word learning. Their basic paradigm was as follows: participants learned invented names for sixteen irregular shapes across four blocks (each of 64 trials). Each trial took the form of two shapes on the screen, one on the left, and one on the right, along with a speech recording of a name. The name belongs to one shape, the *target*. The other shape is a *foil* randomly selected from the other 15 shapes. Participants have to indicate which of the two shapes is the

target. Though at first they have to guess, over time they build up information about the identity of each word's referent through repeated co-occurrence (because foils are chosen randomly and therefore each name appears with the shape it names far more often than with any other shape).

In each of Monaghan, Mattock, and Walker's two experiments, there were two categories of shape, and two categories of names. Eight shapes were rounded, and eight were spiky; likewise eight names were iconically round, while the other eight were iconically spiky (cf. the Köhler, 1929; Ramachandran & Hubbard, 2001). Iconic words were created using phonetic features identified in previous literature as suggesting roundness or spikiness. Monaghan et al.'s Experiment 1 varied consonants: names were created using a one-syllable CVC template, with plosives in the onset and coda positions for the spiky names, and nasals, liquids, or approximants in the onset and coda position for the round words, holding a range of vowels constant. Experiment 2 used words of the same form, contrasting back vowels in the round condition with front vowels in the spiky condition, holding a range of consonants constant.

Assignment of names to shapes could either be *congruent* or *incongruent*. Congruent pairings mapped a name with an iconically fitting shape (i.e. round name-round shape; or spiky name-spiky shape). Incongruent pairings did the opposite (round name-spiky shape, spiky name-rounded shape). For each participant, half of the shapes and half of the names in each category received congruent pairings, and the other half received incongruent pairings.

The primary question was whether iconically congruent names would be easier to learn (i.e. elicit higher accuracy) than incongruent names. Monaghan et al. found that

this was the case, but with interesting caveats. Firstly the advantage of congruence only appears *after* the first block. Furthermore the advantage of congruence was only present in trials where the foil was from the opposite category of shape to the target. In effect, giving e.g. a round shape a round name improves performance when the foil is a spiky shape (i.e. where the foil would be an incongruent pairing with the name), but not when the foil is another round shape. From the first point Monaghan et al. conclude that the congruence advantage is not merely the result of referential disambiguation: i.e. participants showing a bias towards iconic guesses where they are unsure of the referent of the word (otherwise it should be there from the beginning). Rather *bona fide* learning – i.e. extraction and encoding of information from past encounters with names – proceeds faster for congruent pairings than for incongruent pairings. From the second result Monaghan et al. argue that only some aspects of learning are facilitated by iconicity. Specifically, the category of the shape mapped to (i.e. round or spiky) is learned faster for iconic pairings, but the identity of the shape *within* that category is not. This is in keeping with the conclusions of Monaghan, Christiansen, & Fitneva (2011), who argue that systematicity is advantageous between categories, but arbitrariness is optimal within them. However, it is equally consistent with the claim that iconicity is useful at whatever level it can be had, given that in this instance iconicity only applied to category-level properties of the shapes. This may well be the more plausible assumption, given that sign languages incorporate massive amounts of iconicity, and this iconicity is typically concept- rather than superordinate category-specific. e.g. The BSL signs for RABBIT, CAT, GORILLA etc. incorporate iconicity for rabbits, cats, and gorillas, not for animals generally.

In particular, the absence of a congruence advantage in the first block in Monaghan et al. is striking. This is because it seems to contradict a prominent set of results in the field: Köhler and his successors' *takete-baluba/kiki-bouba* experiments (Bremner, Caparos, Davidoff, de Fockert, Linell, & Spence, 2012; Köhler, 1929; Maurer, Pathman, & Mondloch, 2006; Ramachandran & Hubbard, 2001, see Chapter 2 for a review). The upshot of these experiments is that when forced to guess, participants highly reliably pair iconically spiky words with spiky shapes, and iconically round words with round shapes. This seems to be close to the situation participants find themselves facing in the first block of Monaghan et al.'s cross-situational learning experiment, before they know any names. Why therefore wouldn't they guess iconically? True, only one word is presented at a time, and it's possible that in the classic kiki-bouba study the iconic properties of the names might be heightened by their mutual contrast. Moreover, unlike Köhler-style studies, cross-situational learning is framed to participants as a learning rather than a guessing task. Nonetheless, it is surprising that no bias is evident. It is also theoretically important, given that a bias would indicate that the learning advantage conferred by iconic vocabulary could be down to referential disambiguation. We will return to this question when discussing the results of Experiment 1, a near replication of Monaghan et al. (2012), where I obtain subtly but importantly different results.

Beyond Monaghan et al., who show that sound-shape iconicity improves performance in cross-situational learning, the paradigm has been little used to investigate iconicity. In Experiments 2-4 I use the paradigm to investigate the mechanism of sound-shape iconicity for the first time, testing whether the round and spiky effects are equally strong.

# Experiments

## **Experiment 1: Replication of Monaghan, Mattock, and Walker (2012)**

Here we perform a near replication of Monaghan, Mattock, and Walker (2012), both in order to replicate their results, and in order to baseline our cross-situational learning paradigm for further experiments. The principle difference with Monaghan et al. was that we constructed our words on the basis of norms rather than phonetic features, for reasons discussed below.

### **Methods**

**Participants** Twenty four adult native English monolinguals (13 women,  $M = 29.7 \pm 10.0$ ) were recruited from the UCL SONA subject pool to participate in an ‘implicit learning of names for shapes’ experiment, in exchange for cash or course credit. Seven participants who failed to learn (i.e. to achieve above-chance accuracy on the last block at  $p \leq .05$ ) were excluded and replaced.

**Visual stimuli (shapes)** Sixteen shapes were created using the GNU Image Manipulation Program. Initially eight ‘spiky’ shapes (irregular four pointed stars) were created using randomised parameters. Eight ‘rounded’ shapes were created by taking each spiky shape and using its corners as fixed points for Bezier curves describing a bulbous form, and then scaling to match for size (see Figure 3.1 for example shapes). Stimuli were 600\*600 pixels images comprising the shape in black on a white background.

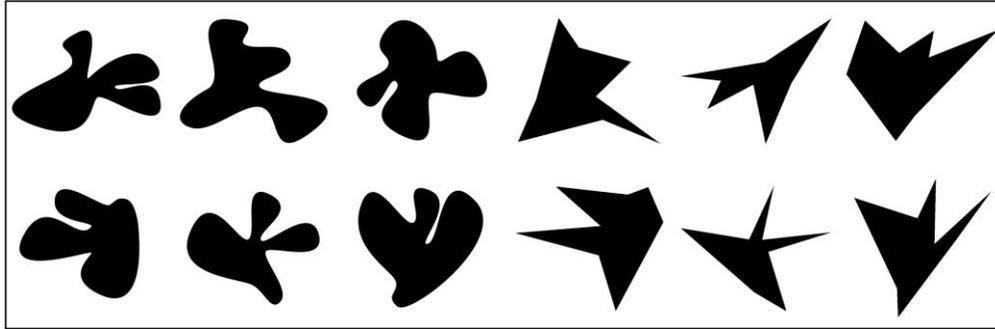


Figure 3.1: Examples of round and spiky shapes used in Experiment 1

**Auditory stimuli (names)** Names were constructed on the basis of *LetterScore*, a text-based index of sound-shape iconicity. Normative data were obtained in the following way: All consonant-vowel pairings possible in English orthography that feature consonants with only one canonical pronunciation (N=85; c, g, q, and x were excluded) were rated by monolingual English speakers who did not participate in the other studies (N = 28, 12 women,  $28.5 \pm 12.0$  years old) on a ten-point scale anchored by a circle (1) and a star (10) (see Appendix 3.1). A centred scale was created by redefining the mean rating (5.04) as zero (see Appendix 3.2 for a full list of syllables).

Eight of the names were constructed using syllables that received the spikiest ratings, eight using the syllables that received the roundest ratings. Within each category of name, two were one syllable long, four were two syllables long, and two were three syllables long (see Table 3.1). Syllables were assigned to words randomly, and concatenated randomly within polysyllabic words. The recordings of the names used in the experiment were made by a female native speaker of North American English, pronouncing the written words as she considered natural (see Table 3.1 for IPA transcriptions).

Subsequently the word recordings were normed as part of a wider word norming study. 101 native English speakers ( $M = 32.4 \pm 9.7$ , 41 women) recruited through the website Prolific Academic were each given 118 speech tokens to rate (largely from another study), meaning that each of the 1180 speech tokens from the experiment was rated about ten times on average. The study was performed online using Qualtrics (2015). In each trial, the participant saw a seven-point ratings scale. For half of participants 1 was anchored by a round shape and 7 by a spiky shape, for the other half this was reversed. After ratings were collected those collected using the round-high scale were ‘flipped’ such that 7 became 1, and 1 became 7 etc.. This meant that for all data ‘1’ represented the roundest rating, and ‘7’ the spikiest. The mean of each token’s ratings was then taken. This was its WordScore (see Table 3.1).

<b>Shape Category</b>	<b>Word</b>	<b>Pronunciation</b>	<b>Mean Rating</b>
Spiky	ka	[kɑ]	4.91
Spiky	kevi	[ 'kɛvi]	3.55
Spiky	pepi	[ 'pɛpi]	4.75
Spiky	ta	[tɑ]	5.00
Spiky	tikiza	[tɪ'kizə]	4.91
Spiky	zefi	[ 'zɛfi]	5.00
Spiky	zidiyi	[zidi'ji]	4.40
Spiky	zuji	[ 'zudʒi]	5.20
Round	bo	[bəʊ]	2.91
Round	hodusu	[həʊdu'su]	3.38
Round	lu	[lu]	3.50
Round	majuwu	[mɑ'dʒuwu]	3.00
Round	mebu	[mɛ'bu]	2.73
Round	muyo	[ 'mujəʊ]	2.11
Round	woso	[ 'wəʊsəʊ]	2.90
Round	yulo	[ 'juləʊ]	2.70

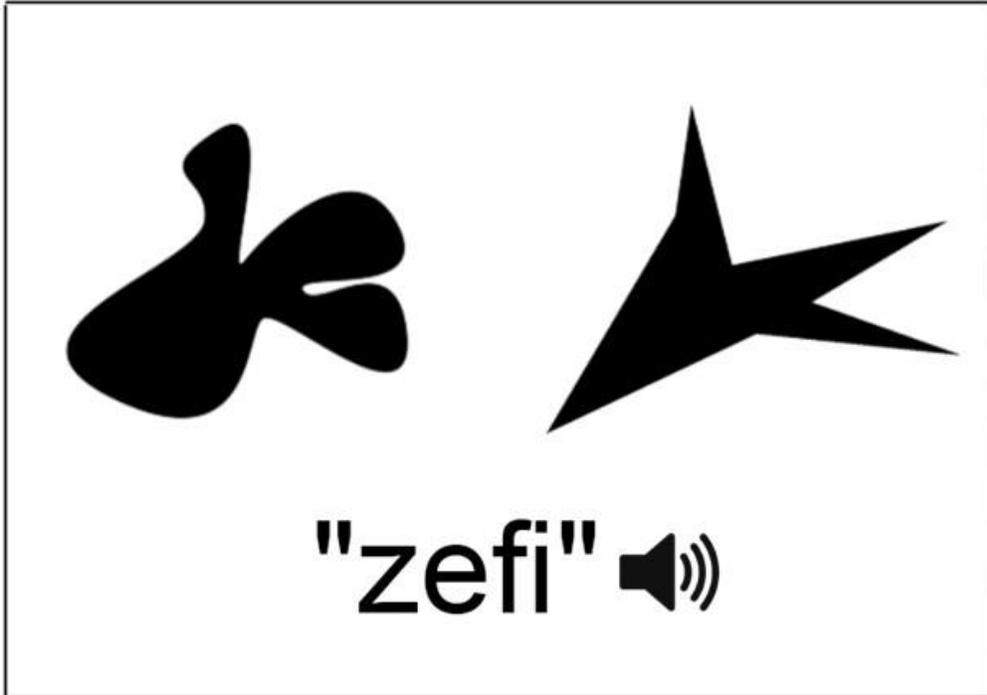
*Table 3.1: WordScores for stimuli in Experiment 1*

T-tests confirm that the spiky names ( $M = 4.71 \pm 0.53$ ) were rated as significantly spikier than the round names ( $M = 2.90 \pm 0.43$ ) ( $p < .001$ ,  $t(13.4) = 7.54$ , difference = 1.81, 95% CI [1.29, 2.33]; Cohen's  $d = 3.77$ ).

**Apparatus and Procedure** – The study was run using Matlab 7.4.0 on an IBM compatible PC equipped with a 15" monitor (resolution: 1024×768).

For each participant, each shape was assigned a name. Half of the shapes in each category were assigned *congruent* names (i.e. round names for round shapes, spiky names for spiky shapes). The other half of each category was assigned *incongruent* names (i.e. spiky names for round shapes and round names for spiky shapes). Assignment of names to shapes, and assignment of both names and shapes to incongruent versus congruent pairings were counterbalanced between participants.

The experiment took the form of a series of 256 trials, each featuring two shapes on screen (one to the left and one to the right – see Figure 3.2) and one name (played through headphones). The name belonged to one of the two shapes (this shape was the *target*, the unnamed shape being the *foil*) and the participant's task was simply to say which shape the name belonged to (by pressing the left arrow or the right arrow). Participants did not receive feedback and had to guess at first. However, over time it was possible to infer which name belonged to which shape because each name only consistently appeared with a single shape.



*Figure 3.2: A cross-situational learning trial (note that names were presented aurally, not in text)*

Trials were grouped into four blocks of 64 trials each. Within each block each name appeared four times, and concomitantly each shape appeared four times as a target and four times as a foil. The number of times each shape appeared on each side of the screen in each role was counterbalanced, as was the number of appearances by each shape as a foil for a target from its own category vs. the opposite category. The same name was not permitted to appear for two trials in a row. Within these constraints, trials and trial order were randomised.

## **Results**

Trials with reaction times of less than 0.5 seconds or more than 25 seconds were removed in this and all other experiments in the chapter.

Data was analysed using logistic mixed effects regression using the package lme4, version 1.1-12 (Bates, Maechler, Bolker, & Walker, 2015) running in R version 3.2.1 (R Core Team, 2015). In addition to random intercepts for names and participants, we also included random slopes. We aimed for a design-driven maximal random effects structure (see Barr, Levy, Scheepers, & Tily, 2013), but as the number of observations per statistical unit in the random effects structure was comparatively small, and the fitting of binomial models such as these is computationally intensive, we found that we were limited in the number of random effects we could fit. For participants we therefore included random slopes for linear block, congruence, category of foil (coded as same or different to category of target), and the interaction between congruence and category of foil (which proved important in Monaghan, Mattock, and Walker, 2012). For names, we were limited to random effects slopes for congruence, category of foil, and their interaction. Within each of the two random effects grouping variables we allowed unstructured covariance matrices.

Block was coded linearly (1 = -1.5, 2 = -0.5, 3 = 0.5, 4 = 1.5), and both other variables were contrast coded (incongruent = -0.5, congruent = 0.5; same category foil = -0.5, different category foil = 0.5; see Table 3.2).

Variable	Level	Contrast Code	Shorthand Name
<i>Block</i>	First Block	- 1.5	<u>B1</u>
<i>Block</i>	Second Block	- 0.5	<u>B2</u>
<i>Block</i>	Third Block	+ 0.5	<u>B3</u>
<i>Block</i>	Fourth Block	+ 1.5	<u>B4</u>
<i>Congruence</i>	<i>Incongruent</i> pairing of name and shape	- 0.5	<u>-Congruent</u>
<i>Congruence</i>	<i>Congruent</i> pairing of name and shape	+ 0.5	<u>+Congruent</u>
<i>Category of Foil</i>	Foil shape from <i>same</i> category as target	- 0.5	<u>-Different</u>
<i>Category of Foil</i>	Foil shape from <i>different</i> category to target	+ 0.5	<u>+Different</u>

Table 3.2: variable codes for Experiment 1

Our procedure was to construct an initial omnibus model featuring all predictors and all interactions, to examine the results of this omnibus model, and then to construct a final model from which we have removed the predictor of quadratic generation if it is unreliable, and any unreliable interactions (unless they are components of reliable interactions - i.e. a reliable three-way interactions means keeping each of the two-way interactions obtained by removing one of its predictors).

Our dependent variable was accuracy: i.e. whether participants answered correctly on given trials.

**Omnibus Model** Our first omnibus model featured fixed effects of congruence, category of foil (again coded as same or different to target), and linear and quadratic polynomials of linearly coded block, as well as a fixed intercept (see Appendix 3.3 for specification). All interactions up to the full three-way interaction were included (N.B. linear and quadratic block were never included in the same interaction). In this first

model the only reliable interaction was between congruence and category of foil - all other interactions were unreliable ( $|z| < 1.5$ ) and were removed from the next model.

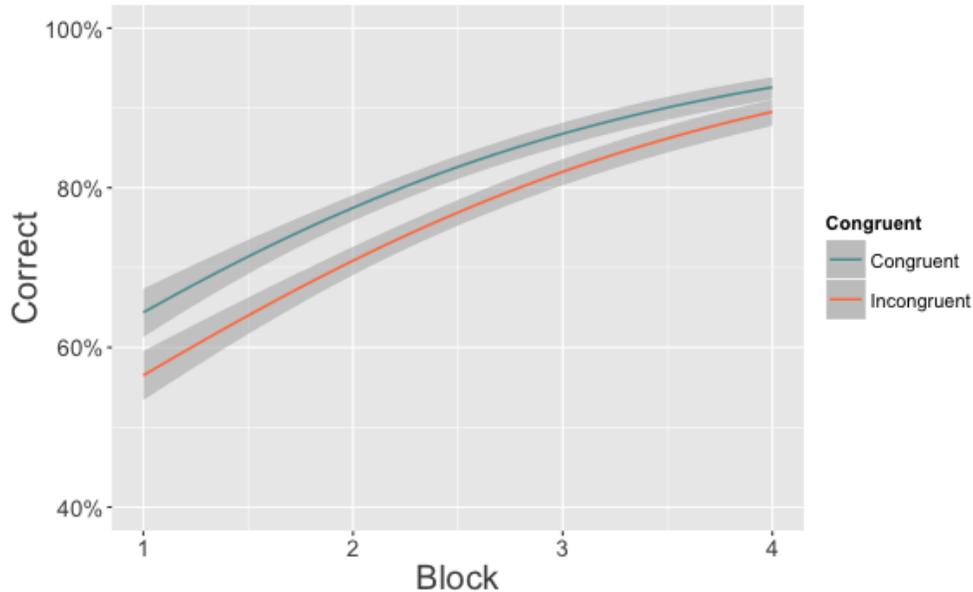


Figure 3.3: Graph of the predictions by block and congruence of the final omnibus model for Experiment 1<sup>18</sup>. Error bars represent 95% confidence intervals.

This new reduced model showed reliable effects of linear block<sup>19</sup> ( $\beta = 0.832$ , 95% CI [0.655, 1.010],  $z = 9.188$ ), indicating that participants improved over the blocks, i.e. learned; and of congruence ( $\beta = 0.3604$ , 95% CI [0.131, 0.590],  $z = 3.074$ ), indicating that overall, participants performed better in trials where they heard a congruent name (see Figure 3.3). The interaction between congruence and category of foil was also significant ( $\beta = 0.700$ , 95% CI [0.3414, 1.059],  $z = 3.826$ ), indicating

<sup>18</sup> Note that here block is plotted as a continuous variable, whereas in fact it was ordinal. This reflects the fact that in the model the codes for block were numerical rather than categorical. However to see the predictions for each block we can simply look at the slice defined by the relevant point on the x-axis (1, 2, 3, or 4). The same will hold for the graphs of the predictions of the other models.

<sup>19</sup> N.B. parameters are expressed in terms of log odds ratios, and are therefore not immediately interpretable in terms of e.g. group means. However, if a significant coefficient is positive it can be interpreted as indicating that performance tends to be better in the group with a positive contrast code for that variable. A negative coefficient implies the opposite.

that congruence represents more of an advantage in trials where the foil shape is from the opposite category to the target.

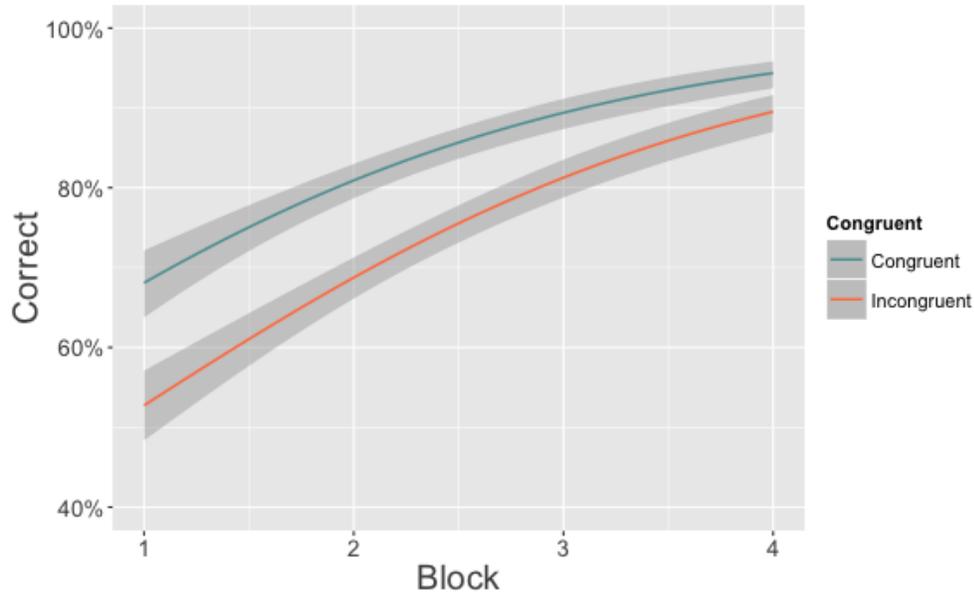


Figure 3.4: Graph of the predictions by block and congruence of the +Different model for Experiment 1. Error bars represent 95% confidence intervals.

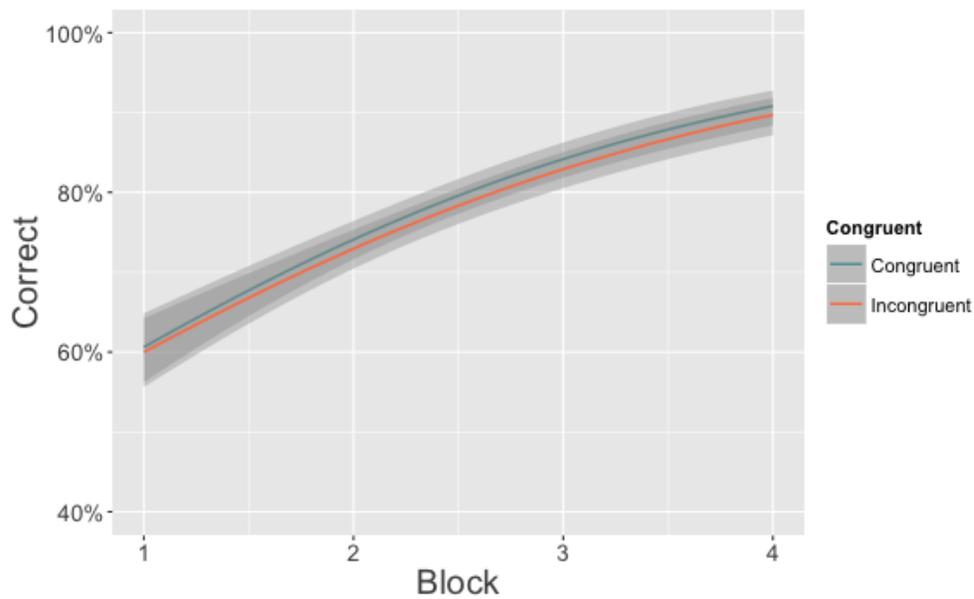


Figure 3.5: Graph of the predictions by block and congruence of the -Different model for Experiment 1. Error bars represent 95% confidence intervals.

**Category of Foil** In order to check how much of an advantage congruence conferred in trials with different categories of target and foil (+Different; see Table 3.2) vs. trials with the same category of target and foil (-Different), we ran separate models for each. Interactions that were not significant in the omnibus model were omitted, but random effects were kept the same (except that slopes for category of foil were removed, of necessity). +Different trials showed the expected effect of linear block ( $\beta = 0.907$ , 95% CI [0.708, 1.105],  $z = 8.957$ ), and – crucially – a reliable effect of congruence ( $\beta = 0.718$ , 95% CI [0.398, 1.037],  $z = 4.402$ ), indicating that in +Different trials performance tended to be better for congruently named stimuli (see Figure 3.4). However, though linear block was also reliable in the model for – Different trials ( $\beta = 0.814$ , 95% CI [0.619, 1.010],  $z = 8.164$ ), congruence was not ( $\beta = -0.030$ , 95% CI [-0.276, 0.216],  $z = -0.240$ ). See Figure 3.5. Overall, these results indicate that as in Monaghan et al.'s original study, congruence was only an advantage where the foil and the target were from different categories. As discussed earlier, this is unsurprising.

**Block 1** Monaghan et al. found that congruence interacted with block. Crucially for them, there was no congruence advantage in the first block, indicating that the benefit of congruence was to learning proper rather than merely reflecting a bias to respond to iconically congruent stimuli. By contrast, we found no interaction between congruence and block, suggesting that a congruence advantage was present throughout the experiment, including the first block. To test this, we fitted a model for the first block only, again excluding interactions not present in the omnibus model and retaining random effects (except slopes for block). There were reliable effects of congruence ( $\beta = 0.328$ , 95% CI [0.047, 0.609],  $z = 2.288$ ) – indicating that

performance was better in congruent trials – and of the interaction between congruence and category of foil ( $\beta = 0.842$ , 95% CI [0.368, 1.316],  $z = 3.484$ ), indicating that the benefit of congruence was stronger in +Different trials, just as in the experiment as a whole.

However, though this result is consistent with the iconicity advantage being the result of bias rather than learning, it is also consistent with the contrary, as long as learning took place during the first block. Therefore we examined the trials where names appear for the first time. Excluding –Different trials (where neither the learning nor the bias account predicts an iconicity advantage), we focused on +Different trials. Here the learning account predicts that there should be no iconicity advantage (as the name has not yet been encountered), but the bias account predicts the same iconicity advantage seen throughout the experiment.

I took the 187 +Different trials where a participant encountered a name for the first time and fitted a simple binomial linear mixed effects model, featuring a fixed intercept, and random intercepts by participant. On 56.1% of trials participants chose the iconically congruent referent for the name, choosing the mismatching referent 43.9% of the time. The model's intercept was not reliably different from zero<sup>20</sup> under two-tailed interpretation ( $\beta = 0.247$ , 95% CI [-0.048, 0.549],  $z = 1.678$ ), but under a one-tailed interpretation predicting that any deviation from zero should be in the direction of iconic congruence, the intercept is significantly different from zero at  $p = .047$ ). Thus though this analysis has lower power than we would like, it suggests that a bias towards iconic matches is present before learning has taken place.

---

<sup>20</sup> Zero here would indicate that participants were at chance. Values of greater than zero indicate that participants were more likely to chose the iconically congruent match.

In conclusion, we largely replicated Monaghan, Mattock, and Walker's (2012) findings, the difference being that we found an advantage of iconicity from the first block. This difference with Monaghan et al. – taken with the fact that iconic congruence is only an advantage in trials where the foil is from the opposite category – is consistent with the possibility that iconicity did not enhance learning per se, but rather only biased participants towards the right answer in trials where they were forced to guess, effectively assisting with *gavagai*-style disambiguation. That such a bias should be in play is unsurprising given the previous “kiki-bouba” literature on name guessing. The discrepancy with Monaghan et al. may be related to differences in name stimuli: while we tailored ours to maximize iconicity, they created theirs on the basis of phonetic features. The members of each these feature classes may tend to connote either roundness or spikiness, but there may also be exceptions within them (e.g. Monaghan et al. used plosives as spiky sounds, but [b] – a plosive – is widely deemed to sound round, cf. *bouba*). Perhaps the comparative subtlety of Monaghan et al.'s iconicity required a block or so of exposure before participants became aware of it, whereas ours was apparent immediately.

Next we move on to a set of experiments that represent a variation on Experiment 1, aimed at testing the relative contribution of roundness and spikiness to sound-shape iconicity.

## **Experiment 2**

Experiment 1 replicated Monaghan et al.'s (2012) result that cross-situational learning performance is better when stimuli have iconically congruent names. However, as discussed in the last two chapters, the mechanisms of sound-shape

iconicity remain quite unclear. Experiment 2 aims to provide data to help clarify this situation by modifying Experiment 1 in order to test the effect of round-to-round and spiky-to-spiky iconicity separately. This is achieved by using iconically neutral names as well as round and spiky names. The basic idea is to see how much of an advantage round names for round shapes enjoy over neutral names for round shapes, and how much of an advantage spiky names for spiky shapes enjoy over neutral names for neutral shapes.

At first sight it might seem that a comparison of the strength of round and spiky iconicity could be achieved using results from Experiment 1, simply by asking whether *both* round and spiky shapes enjoy an iconicity advantage when paired with a name from their respective categories. However, suppose that e.g. round iconicity is stronger, with spiky iconicity weaker or absent. Even under these circumstances spiky names might be preferentially paired with spiky shapes on the basis of a principle of contrast (“shapes that contrast with round shapes should be paired with words that contrast with round words”), or a bias towards choosing each category of shape 50% of the time (otherwise a bias towards choosing round referents for round shapes and no bias regarding spiky shapes would lead to round shapes being chosen more than 50% of the time). Similar considerations could explain how classic sound-shape iconicity experiments like the kiki-bouba paradigm, or norming studies, have thus far masked an asymmetry.

Simply adding neutral names to the one-condition format used in Experiment 1 poses the same problems. Therefore we opted for a two-condition design. Each condition is of the same format as Experiment 1, and each features both round and spiky shapes, but one condition features round and neutral names only, the other features spiky and neutral names only. In this scenario there is no potential for one

class of name (round or spiky) to “lend” the other iconicity by contrast, because the two never appear together. If one class of name is less iconic (or not really iconic at all) then we would expect minimal benefit of one class of shape being paired with that class of name vs. a neutral name.

## **Methods**

**Participants** were 32 adult native English monolinguals (17 women,  $M = 23.3 \pm 4.4$ ), recruited from the UCL SONA subject pool to participate in an ‘implicit learning of names for shapes’ experiment in exchange for cash or course credit. Six participants who failed to perform significantly above chance (at  $p \leq .05$ ) in one or both of their final blocks were excluded and replaced.

## **Materials**

**Visual Stimuli (Shapes)** In addition to the eight round and eight spiky shapes used in Experiment 1, an additional eight of each were created in the same manner (see Figure 3.1 for examples).

**Auditory stimuli (names)** Thirty two names were generated using previously normed syllables (see Experiment 1) - eight composed of syllables normed as round, eight of syllables normed as spiky, and sixteen of syllables normed as neutral. The round and spiky names were the same as those in Experiment 1. Names were recorded by a female native speaker of North American English (see Table 3.3 for transcriptions).

Shape Category	Word	Pronunciation	Mean Rating
Spiky	ka	[kɑ]	4.91
Spiky	kevi	[ 'kɛvi]	3.55
Spiky	pepi	[ 'pɛpi]	4.75
Spiky	ta	[tɑ]	5.00
Spiky	tikiza	[tr'kizə]	4.91
Spiky	zefi	[ 'zɛfi]	5.00
Spiky	zidiyi	[zidi'ji]	4.40
Spiky	zuji	[ 'zudʒi]	5.20
Round	bo	[bəʊ]	2.91
Round	hodusu	[həʊdu'su]	3.38
Round	lu	[lu]	3.50
Round	majuwu	[mɑ'dʒuwu]	3.00
Round	mebu	[mɛ'bu]	2.73
Round	muyo	[ 'mujəʊ]	2.11
Round	woso	[ 'wəʊsəʊ]	2.90
Round	yulo	[ 'juləʊ]	2.70
Neutral	dabiye	[ 'dɒbɑɪji]	4.60
Neutral	fasi	[ 'fɑsɑɪ]	3.00
Neutral	fu	[fu]	2.56
Neutral	jahe	[ 'dʒɑhi]	4.56
Neutral	je	[dʒi]	4.78
Neutral	kosere	[kəʊsi'ri]	3.80
Neutral	naha	[ 'nɑhɑ]	3.43
Neutral	nebetu	[ni'bitu]	4.89
Neutral	puhi	[ 'puhɑɪ]	2.86
Neutral	ravu	[rɑ'vu]	4.00
Neutral	rumiya	[ru'mɑɪjɑ]	2.83
Neutral	sato	[ 'sɑtəʊ]	3.90
Neutral	va	[vɑ]	3.50
Neutral	vonu	[ 'vəʊnu]	3.33
Neutral	wa	[wɑ]	3.30
Neutral	wewi	[ 'wiwɑɪ]	5.00

Table 3.3: WordScores for stimuli in Experiment 2

Names were rated as part of the same norming study described in the “Auditory Stimuli” section of Experiment 1. T-tests confirm that the spiky names ( $M = 4.71 \pm 0.53$ ) were rated as significantly spikier than the round names ( $M = 2.90 \pm 0.43$ ) ( $p <$

.001,  $t(13.4) = 7.54$ , difference = 1.81, 95% CI [1.29, 2.33]; Cohen's  $d = 3.77$ ). Moreover, neutral names ( $M = 3.77 \pm 0.80$ ) were rated as significantly less spiky than spiky names ( $p = .002$ ,  $t(20.0) = 3.46$ , difference = 0.94, 95% CI [0.37, 1.51]; Cohen's  $d = 1.31$ ), and significantly less round than round names ( $p = .002$ ,  $t(21.8) = 3.46$ , difference = 1.04, 95% CI [0.35, 1.39]; Cohen's  $d = 1.24$ ).

**Apparatus and Procedure** Every participant took part in two conditions, each featuring separate names and shapes. Each condition was of identical form to Experiment 1, including counterbalancing and prohibition on repeated names.

One of the two conditions was the 'round' condition. In this condition half of the shapes were round and half spiky (eight of each), and, crucially, half of the *names* were round and half *neutral*. The other condition was the 'spiky' condition – which again had eight round and eight spiky shapes (different to the ones used in the round condition), but by contrast had eight neutral names and eight *spiky* names (again, neutral names were new). Shapes and neutral names were counterbalanced across conditions between participants. Condition order was also counterbalanced between participants.

At the outset of each participant's experiment, each shape was assigned a name from its condition (specific assignment was counterbalanced between participants). Half of the shapes in each category in each condition were assigned iconically congruent names. The other half of each category were assigned iconically incongruent names.

Here congruence is defined *within* whichever half of the putative round-spiky spectrum of sounds the condition in question covers. In the round condition, round

name-round shape pairings were considered congruent and round name-spiky shape pairings were considered incongruent. However, neutral name-spiky shape pairings were considered congruent for the purposes of the following analysis (as there are no spiky names in this condition, and also in contrast to the incongruent round-name-spiky shape pairings) and neutral name-round shape pairings were considered incongruent (as they are less congruent than round-round pairings). The converse applied for the spiky condition (see Table 3.4). This condition-relative way of thinking of congruence may seem clearer if we imagine that one of the sound-shape associations, say the round-round association, really is primary. In that case, all along iconic congruence has only been a matter of whether a name/shape has been round, vs. whether it has been anything else. A rounded shape-neutral name pairing really is incongruent (inasmuch as a round shape can receive an incongruent pairing), and a spiky shape-neutral name pairing really is congruent (inasmuch as a spiky name can receive a congruent pairing).

	<b>Round Condition</b>		<b>Spiky Condition</b>	
	<i>Round Name</i>	<i>Neutral Name</i>	<i>Neutral Name</i>	<i>Spiky Name</i>
<i>Round Shape</i>	Congruent	Incongruent	Congruent	Incongruent
<i>Spiky Shape</i>	Incongruent	Congruent	Incongruent	Congruent

*Table 3.4: Relative Congruence*

Terminology aside, the crucial point is that if it is indeed the case that only one of the two iconic pairings is actually effective, then congruence will only be an advantage in that condition. However if both are effective then the conditions will be symmetrical, and it will be an advantage in both conditions.

## Results

Data were analysed using binomial linear mixed effects models in the lme4 package in R. Predictors were condition, block, congruence, and category of foil. Two- and three-way interactions were also included in the model as predictors. As with Experiment 1, computational power imposed limits on the number of random effects we were able to include. Therefore, though we included random intercepts for both names and subjects, random slopes for all main effects by participant, and random slopes for generation, congruence, and category of foil by name (previous models that had not converged properly showed very low variance for the congruence and condition slopes), we were forced to restrict the covariance structure, and omit interactions (see Appendix 3.4). Variables were coded as in Experiment 1. The additional variable of condition was coded as Round = -0.5, Spiky = +0.5.

The initial model failed to converge properly. The variance for the random slope for congruence by name turned out to be very low, so this slope was removed from the next model. In the next model (outlined in Appendix 3.4), the only reliable interaction was between congruence and category of foil; all other interactions were unreliable ( $|z| < 1.5$ ). The interaction between condition and congruence was only marginally reliable ( $\beta = -0.208$ , 95% CI [-0.482, 0.066],  $z = -1.486$ ), and therefore excluded from the stripped model, and the three-way interaction between condition, congruence,

and category of foil was unreliable ( $\beta = -0.047$ , 95% CI [-0.377, 0.283],  $z = -0.277$ ). This latter would have indicated that condition modulates the congruence-by-category of foil interaction seen in Experiment 1.

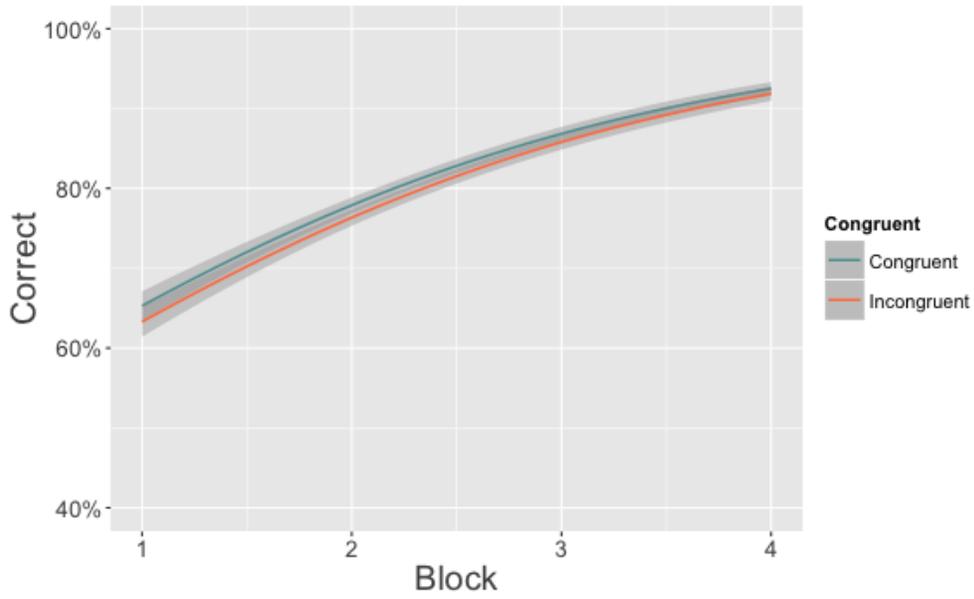


Figure 3.6: Graph of the predictions by block and congruence of the stripped omnibus model for Experiment 2. Error bars represent 95% confidence intervals.

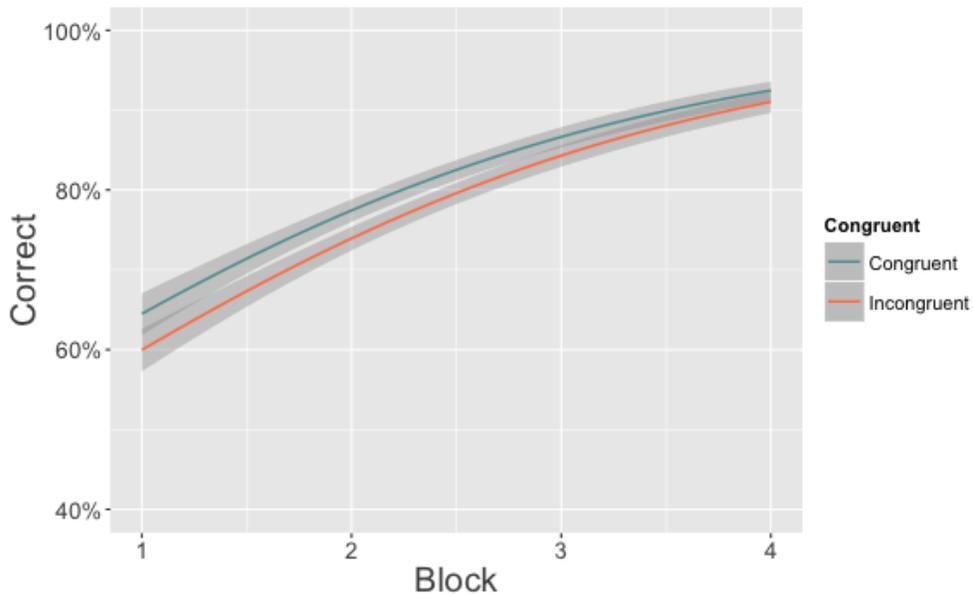


Figure 3.7: Graph of the predictions by block and congruence of a model for the rounded condition only of Experiment 2. The same predictors are included as for the striped omnibus model. Error bars represent 95% confidence intervals.

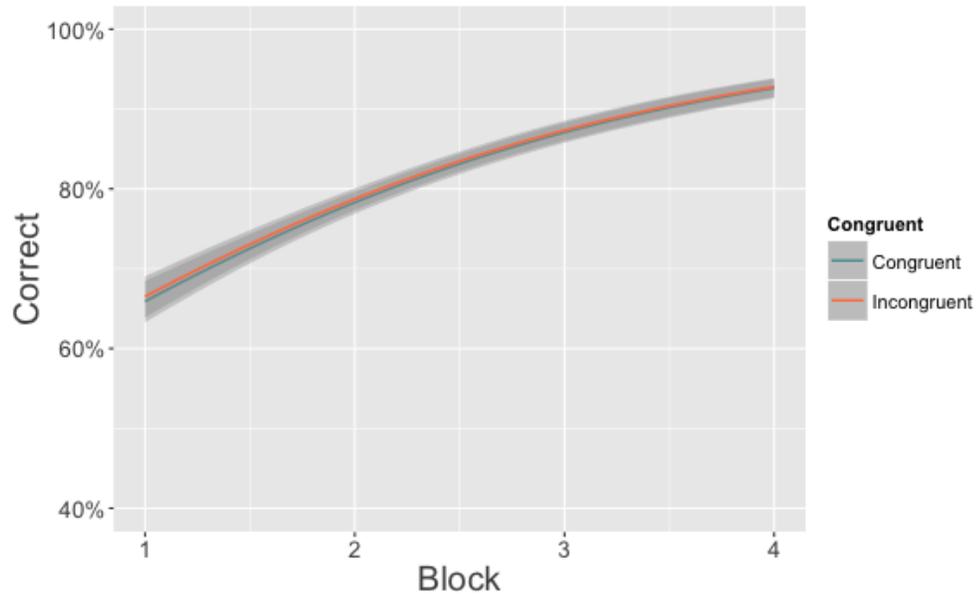


Figure 3.8: Graph of the predictions by block and congruence of a model for the spiky condition only of Experiment 2. The same predictors are included as for the striped omnibus model. Error bars represent 95% confidence intervals.

In a stripped model, both linear ( $\beta = 0.723$ , 95% CI [0.621, 0.826],  $z = 13.868$ ) and quadratic block ( $\beta = -0.109$ , 95% CI [-0.178, -0.04],  $z = -3.104$ ) were reliable predictors, indicating that performance improved over the blocks, with improvement being faster between early than late blocks (see Figure 3.6). Category of foil was also a reliable predictor ( $\beta = 0.167$ , 95% CI [0.068, 0.266],  $z = 3.303$ ), indicating that performance was better on trials with foils from the opposite category to the target. Finally, the interaction between congruence and category of foil was reliable ( $\beta = 0.325$ , 95% CI [0.16, 0.49],  $z = 3.857$ ), reflecting the fact that performance was better on congruent trials as long as the target and foil were from different categories.

Thus, to summarise, Experiment 2 largely replicated the result of Experiment 1, in that participants improved across blocks, and there was an advantage of congruence in +Different trials. It did not find an unambiguously reliable difference between round and spiky conditions in terms of how big an advantage iconic congruence conferred in each case (see Figures 3.7 and 3.8 for a comparison). However, in light of the marginally reliable interaction noted above, this could be due to limited statistical power, as discussed below.

## **Experiments 3 and 4**

Though Experiment 2 failed to find reliable effects of condition, it is possible that this was due to limited power. The interaction between condition and congruence came close to being reliable ( $\beta = -0.208$ , 95% CI [-0.482, 0.066],  $z = -1.486$ ), in the direction that would have suggested that congruence is a bigger advantage in the round condition. This directionality is noteworthy, as it is what would be predicted on the basis of unimodal accounts of the effect, which would claim that it is based on co-occurrence of rounded sounds and rounded lips, without any such experiential basis for spiky iconicity (with presumably comes about via a principle of contrast. Furthermore, it dovetails with results I will report in the next chapter (Chapter 4).

Therefore, we replicated Experiment 2 with the intention of either reproducing or overturning its results on the interaction of condition with congruence.

## **Experiment 3**

### **Methods**

**Participants** were 32 adult native English monolinguals (21 women,  $M = 21.8 \pm 3.2$ ) recruited from the UCL SONA subject pool to participate in an ‘implicit learning of names for shapes’ experiment, in exchange for cash or course credit. Eight participants who failed to perform significantly above chance (at  $p \leq .05$ ) in one or both of their final blocks were excluded and replaced.

## **Materials**

**Visual Stimuli (Shapes)** Were identical to those used in Experiment 2.

**Auditory stimuli (names)** A fresh set of 32 names (eight round, eight spiky, and 16 neutral) were generated using the same method as in Experiment 2. However, an additional factor was controlled this time: in Experiment 2 we simply chose syllables to concatenate on the basis of their rating in the LetterScore norming study, without matching for the number and distribution of phonemes in each category of name. As it happened, spiky names contained a slightly wider range of vowels than round names, and neutral names contained a wider range of consonants than either round or spiky names. There is no reason to assume that this would have confounded the crucial interaction between condition and congruence, as each category of name appears in both congruent and incongruent pairings. However, we decided to control this here in case of any disruptive effect (see Table 3.5 for names). Names were recorded by a female native speaker of North American English, using natural intonation.

<b>Shape Category</b>	<b>Word</b>	<b>Pronunciation</b>	<b>Mean Rating</b>
Spiky	ditipi	[daitaɪ'paɪ]	4.73
Spiky	fi	[faɪ]	4.30
Spiky	kavizu	[kɑ'vaɪzʊ]	5.30
Spiky	niyi	[ 'naɪjɑɪ]	4.89
Spiky	peke	[pi'ki]	5.25
Spiky	ta	[tɑ]	4.20
Spiky	zate	[ 'zɑti]	4.80
Spiky	zeji	[ 'zidʒɑɪ]	4.90
Round	bodū	[ 'bʊdʊ]	3.00
Round	howo	[ 'həʊwəʊ]	2.00
Round	la	[lɑ]	2.50
Round	lowuro	[lʊ'wʊrʊ]	2.82
Round	luso	[ 'lʊsʊ]	3.44
Round	mamu	[mɑ'mʊ]	2.67
Round	mejudo	[mi'dʒʊdʊ]	4.09
Round	yo	[jʊ]	2.67
Neutral	bapasa	[bɑpɑ'sɑ]	4.11
Neutral	biru	[ 'baɪrʊ]	3.38
Neutral	dafa	[dɑ'fɑ]	3.88
Neutral	fe	[fi]	4.00
Neutral	fopo	[ 'fəʊpəʊ]	3.09
Neutral	hisiya	[haɪ'saɪjɑ]	4.00
Neutral	jana	[dʒɑ'nɑ]	5.00
Neutral	jeri	[ 'dʒɪraɪ]	5.00
Neutral	le	[li]	3.45
Neutral	mifu	[ 'maɪfʊ]	3.80
Neutral	netu	[ni'tʊ]	3.75
Neutral	nobeto	[nəʊ'bitəʊ]	3.78
Neutral	nu	[nʊ]	2.40
Neutral	ra	[rɑ]	2.14
Neutral	redevo	[ri'divəʊ]	3.58
Neutral	sehe	[ 'sihi]	5.36

*Table 3.5: WordScores for stimuli in Experiment 3*

Again, names were rated as part of the same norming study described in the “Auditory Stimuli” section of Experiment 1. T-tests confirm that the spiky names ( $M = 4.80 \pm 0.39$ ) were rated as significantly spikier than the round names ( $M = 2.90 \pm 0.63$ ) ( $p < .001$ ,  $t(11.7) = 7.19$ , difference = 1.90, 95% CI [1.32, 2.47]; Cohen’s  $d = 3.60$ ). Moreover, neutral names ( $M = 3.80 \pm 0.86$ ) were rated as significantly less spiky than spiky names ( $p < .001$ ,  $t(22.0) = 3.91$ , difference = 1.00, 95% CI [0.47, 1.53]; Cohen’s  $d = 1.35$ ), and significantly less round than round names ( $p = .01$ ,  $t(18.5) = 2.89$ , difference = 0.90, 95% CI [0.25, 1.55]; Cohen’s  $d = 1.24$ ).

**Apparatus and Procedure** Were as in Experiment 2.

## Results

Data were analysed in the same manner as in Experiment 2 (i.e. binomial LMEM with the same random effects structure – see Appendix 3.5).

The omnibus model featured reliable effect of linear block ( $\beta = 0.739$ , 95% CI [0.589, 0.889],  $z = 9.641$ ), indicating that participants learned over the course of the experiment. However, it featured no significant interactions ( $|z| < 1.6$ ). In this respect, it was different from Experiments 1 and 2, both of which showed an advantage of congruence. However, this may reflect Type II error, as the coefficients for both congruence ( $\beta = 0.177$ , 95% CI [-0.003, 0.357],  $z = 1.929$ ) and the interaction between congruence and category of foil ( $\beta = 0.178$ , 95% CI [-0.081, 0.437],  $z = 1.346$ ) were in the expected direction, with congruence qualifying as marginally reliable. Crucially, the interaction between congruence and condition did not approach reliability ( $\beta = -0.047$ , 95% CI [-0.326, 0.233],  $z = -0.327$ ), and neither did the interaction between congruence, category of foil, and condition ( $\beta = -0.147$ , 95%

CI [-0.464, 0.170],  $z = -0.908$ ) suggesting that inasmuch as there was an effect of congruence, it was no stronger for round names than spiky.

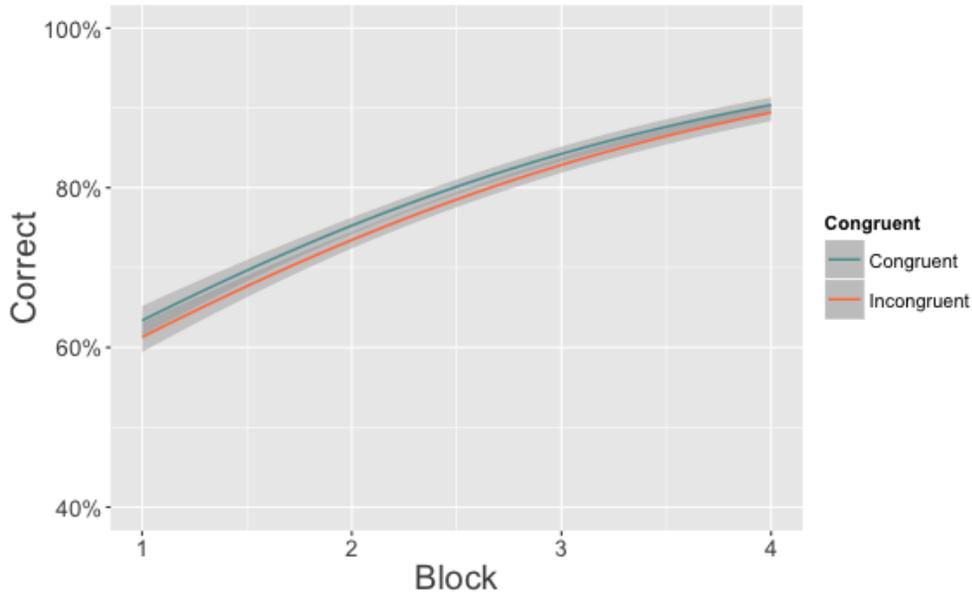


Figure 3.9: Graph of the predictions by block and congruence of the stripped omnibus model for Experiment 3. Error bars represent 95% confidence intervals.

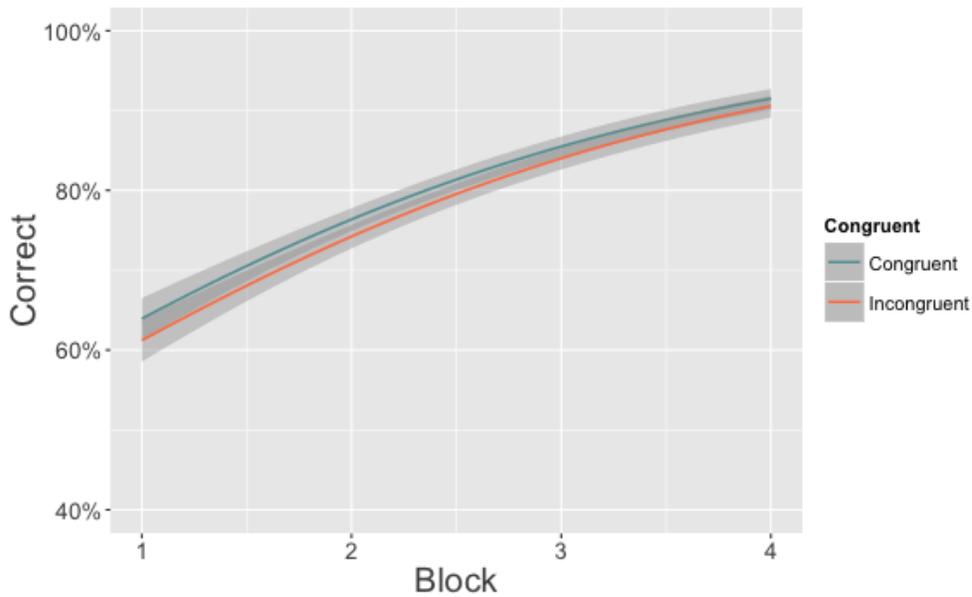
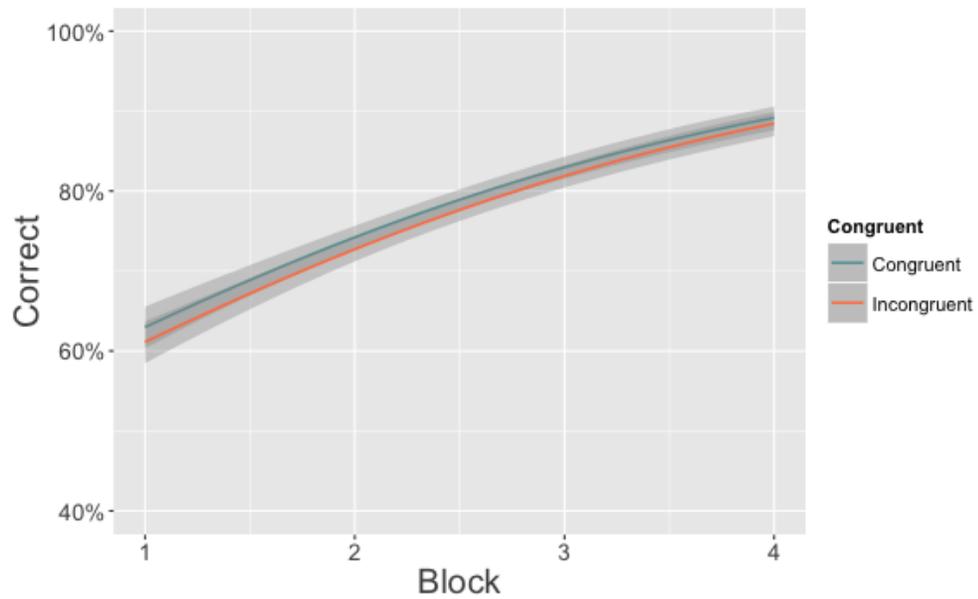


Figure 3.10: Graph of the predictions by block and congruence of a model for the rounded condition only of Experiment 3. The same predictors are included as for the stripped omnibus model. Error bars represent 95% confidence intervals.



*Figure 3.11: Graph of the predictions by block and congruence of a model for the spiky condition only of Experiment 3. The same predictors are included as for the striped omnibus model. Error bars represent 95% confidence intervals.*

The final model was stripped of all interactions, and again showed a reliable effect of linear block only ( $\beta = 0.738$ , 95% CI [0.587, 0.889],  $z = 9.602$ ), again, reflecting learning (see Figure 3.9; see Figures 3.10 and 3.11 for a comparison of the round and spiky conditions).

Experiment 3 again failed to find a reliable interaction between condition and iconic congruence. The temptation is therefore to conclude that none exists, and that therefore neither form of iconicity is more powerful than the other in this kind of task. However, it is well known that inferring the truth of the null hypothesis from a non-significant p-value is a fallacy. Nonsignificance is no guarantee that the null hypothesis is a better explanation for the data than a plausible alternative might be (since null results can be caused by e.g. insufficient statistical power as well as other unknown variables). In order to obtain a more definitive answer to the question, and

after having carried out yet another replication using different stimuli (created following a different strategy), we use Bayesian statistics (Kruschke, 2011).

## **Experiment 4**

### **Methods**

**Participants** were 37 adult native English monolinguals (28 women,  $M = 21.8 \pm 3.2$ ) who took part in the experiment as part of a first-year undergraduate psychology lab. Previously 21 participants were excluded for not performing above chance (at  $p \leq .05$ ) in the last block, and a further 31 were excluded for not being monolingual English speakers (to bring the sample into line with that for the previous experiments).

### **Materials**

**Visual stimuli (shapes)** Were identical to those used in Experiment 2 and 3.

**Auditory stimuli (names)** A fresh set of 24 names (eight round, eight spiky, and 8 neutral – as each participant only participated in one condition) were generated. A slightly different procedure was used to generate the names as compared to earlier experiments. Consonants were selected according to their LetterScores. However vowels were selected according to their frontness vs. backness, a quality that has often been noted as influencing perception of shape (Monaghan, Mattock, & Walker, 2012). Spiky names received front vowels, round names received back vowels, and neutral names received a mixture of both (English has few central vowels). As in

Experiment 3, the range of phonemes was equal for each condition (see Table 3.6). Names were recorded by a female native speaker of North American English, using natural intonation.

<b>Shape Category</b>	<b>Pronunciation</b>	<b>Mean Rating</b>
Spiky	[kɛ]	5.50
Spiky	[ 'kivɛ]	5.14
Spiky	[ 'tɛki]	5.14
Spiky	[tɛ'zivɛ]	4.29
Spiky	[vi'keti]	4.88
Spiky	[ 'vizɛ]	4.91
Spiky	[zi]	5.45
Spiky	[ 'zɛti]	4.90
Round	[bəʊ'luhəʊ]	2.64
Round	[ 'bəʊmu]	2.82
Round	[ 'huləʊ]	2.09
Round	[hu'məʊbu]	2.27
Round	[lu]	2.00
Round	[ 'ləʊbu]	1.67
Round	[məʊ]	3.10
Round	[ 'muhəʊ]	4.11
Neutral	[ 'færeɪ]	2.50
Neutral	[fɑ'jeɪnɑ]	3.27
Neutral	[neɪ'rafeɪ]	3.38
Neutral	[ 'neɪjɑ]	2.56
Neutral	[rɑ]	3.50
Neutral	[ 'reɪnɑ]	2.20
Neutral	[jeɪ]	4.56
Neutral	[ 'jɑfeɪ]	3.40

*Table 3.6: WordScores for stimuli in Experiment 4*

Again, names were rated as part of the later norming study described in the “Auditory Stimuli” section of Experiment 1. T-tests confirm that the spiky names ( $M =$

5.03  $\pm$  0.38) were rated as significantly spikier than the round names ( $M = 2.59 \pm 0.77$ ) ( $p < .001$ ,  $t(10.3) = 7.99$ , difference = 2.44, 95% CI [1.76, 3.12]; Cohen's  $d = 3.60$ ). Moreover, neutral names ( $M = 3.17 \pm 0.75$ ) were rated as significantly less spiky than spiky names ( $p < .001$ ,  $t(10.5) = 6.25$ , difference = 1.86, 95% CI [1.20, 2.51]; Cohen's  $d = 3.13$ ). However, they were numerically but not significantly less round than round names ( $p = .14$ ,  $t(13.9) = 1.53$ , difference = 0.58, 95% CI [-0.23, 1.40]; Cohen's  $d = 0.77$ ).

This lack of a significant difference between round and spiky labels suggests that our method of constructing names for this experiment missed something about how the names are perceived when spoken aloud. This is potentially worrying, as if neutral and spiky names are more distinct than neutral and round names, then this introduces a confound which may increase the effect of congruence in the spiky as compared to round condition. Thus if we were to see effects suggesting that the benefit of congruence is nonetheless stronger in the round condition, this would be strong evidence that round-to-round iconicity is privileged. However if we do not, then interpreting the results may be more problematic. I will nonetheless analyse the results in the same manner as I have for the previous two experiments, and adjust my approach to the Bayesian stats depending on what I find.

**Apparatus and Procedure** Were as in Experiments 2 and 3, except that each participant only took part in one condition (i.e. conditions were between subjects). Also, participants were divided between individual cubicles (as in previous experiments), and a large computer room holding about 40 people.

## Results

Data were analysed using binomial LMEMs as in the previous two experiments. However the random effects structure was modified to reflect the fact that each participant only took part in one condition (see Appendix 3.6).

The omnibus model featured reliable effect of linear block ( $\beta = 0.531$ , 95% CI [0.431, 0.630],  $z = 10.446$ ), indicating that participants learned over the course of the experiment. It also featured a reliable interaction between congruence and category of foil ( $\beta = 0.718$ , 95% CI [0.417, 1.019],  $z = 4.676$ ). It featured no other significant interactions ( $|z| < 1.7$ ). Crucially, the interaction between congruence and condition did not approach reliability ( $\beta = 0.216$ , 95% CI [-0.149, 0.581],  $z = 1.159$ ), and neither did the interaction between congruence, category of foil, and condition ( $\beta = -0.068$ , 95% CI [-0.447, 0.311],  $z = -0.353$ ), suggesting that in as much as there was an effect of congruence, it was no stronger for round names than spiky.

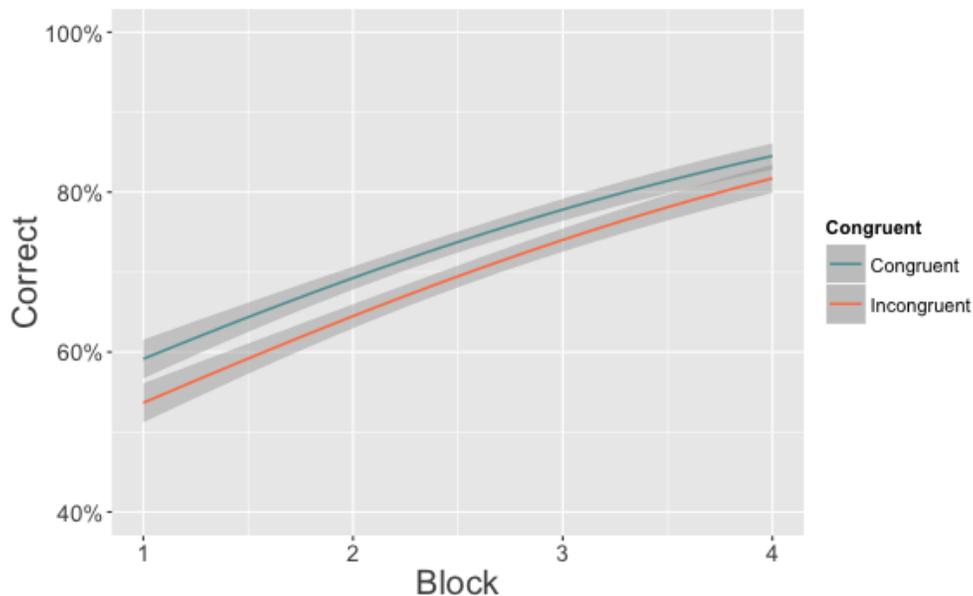


Figure 3.12: Graph of the predictions by block and congruence of the stripped omnibus model for Experiment 4. Error bars represent 95% confidence intervals.

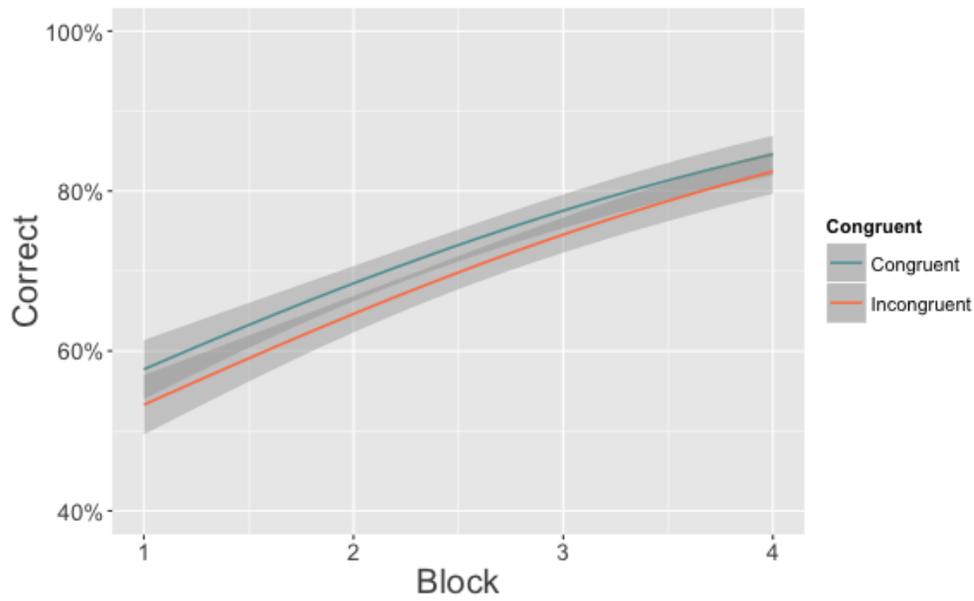


Figure 3.13: Graph of the predictions by block and congruence of a model for the rounded condition only of Experiment 4. The same predictors are included as for the stripped omnibus model. Error bars represent 95% confidence intervals.

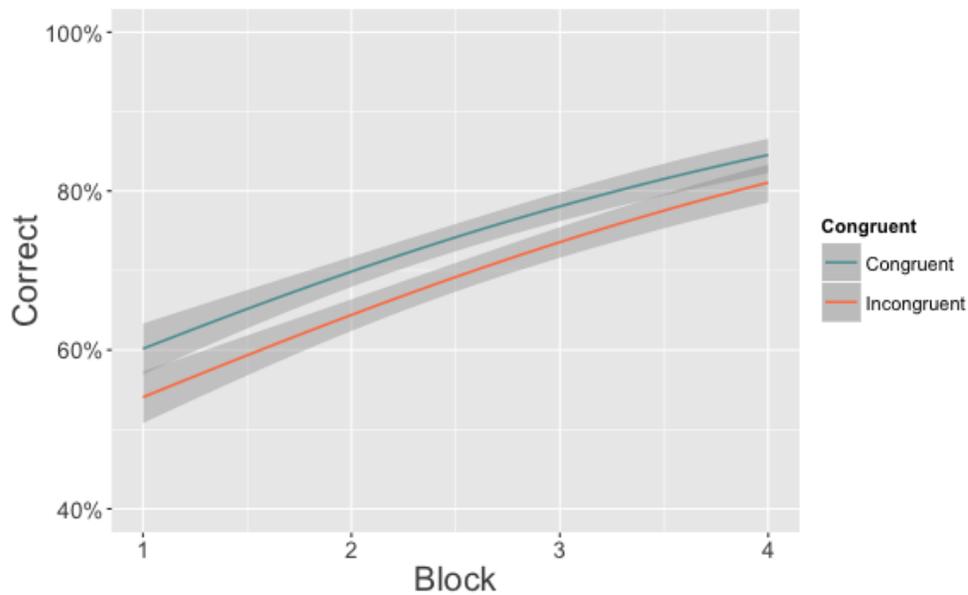


Figure 3.14: Graph of the predictions by block and congruence of a model for the spiky condition only of Experiment 4. The same predictors are included as for the stripped omnibus model. Error bars represent 95% confidence intervals.

The final model was stripped of all interactions except that between congruence and category of foil, and showed reliable effects of linear block ( $\beta = 0.533$ , 95% CI [0.433, 0.632],  $z = 10.535$ ) and congruence ( $\beta = 0.231$ , 95% CI [0.067, 0.396],  $z = 2.762$ ), and a reliable interaction between congruence and category of foil ( $\beta = 0.606$ , 95% CI [0.417, 0.795],  $z = 6.288$ ). Thus again there is evidence for learning, and for an advantage of congruence (see Figure 3.12).

These results are problematic in that though they show no interaction between condition and congruence (see Figures 3.13 and 3.14 for a comparison of conditions), it is possible that a stronger advantage for round-round iconicity is being counteracted by a bigger distinction between categories of name in the spiky condition (see stimuli section). Thus we will perform two Bayesian analyses – one including Experiment 4, and one excluding it.

## **Bayesian Data Analyses of Experiments 2-4**

Increasingly popular in cognitive science, Bayesian statistics are argued to offer a number of important advantages over null hypothesis significance testing - aka NHST (Kruschke, 2011; Morey, Romeijn, & Rouder, 2016; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007).

### **NHST**

NHST means positing a null hypothesis: a particular model of the population which the sample of results that constitutes the dataset is drawn from. The null hypothesis typically reflects the default assumption that across the population some predictor or

set of predictors has no relationship with the dependent variable, i.e. that some parameter value is zero for the population as a whole. As well as specifying relationships between predictors and dependent variables, the null hypothesis also contains an error term – reflecting the fact that the population features variation not captured by the predictors. This can be treated as random (and in parametric statistics is typically assumed to be normally distributed, with variance estimated on the basis of the variance of the sample).

NHST then calls for calculating some statistic from the sample (e.g. the  $t$ - or  $F$ -statistic). These statistics, roughly speaking, give the size of the sample-based estimate of the predictor's effect as compared to the effect specified in the null hypothesis, normalised by the spread of the estimates we would expect under the null merely as a result of random error. From this statistic, in combination with information on the number of independent data points we have, we can calculate a p-value. The p-value represents *the probability that the absolute value of our statistic would be as large as it is, if the null hypothesis were true.*

To take a very simple example, in a two-sample t-test the p-value reflects the probability that the difference between the means of the two conditions would be as great as it is if the null hypothesis that those samples are from the same population were true. Crucially, the only thing that NHST directly tests is how likely the dataset is given a specific model (i.e. the null).

## **Bayesian Statistics**

Bayesian statistics on the other hand explicitly contrasts two or more models (or indeed a continuous space of model parameters), directly testing *the likelihood of the*

*models given the data* (rather than the likelihood of the data given the model, as in NHST).

Bayesian statistics are based on Bayes' theorem, which is derived as follows. We assume a (potentially infinite) set of samples we might observe, and a (potentially infinite) set of possible models under consideration that might describe the populations producing those samples (one of which has to be right one if inference is to work). Start off with the identity:

$$P(M|D)*P(D) = P(D|M)*P(M)$$

where M = model, and D = data. Both sides of the equation simply express the probability<sup>21</sup> P(M&D), i.e. the probability that both a given set of data is observed and the model under consideration is true. This can be rearranged as Bayes' theorem:

$$P(M|D) = P(D|M)*P(M)/P(D)$$

Which implies Bayes' rule:

$$P(M|D) \propto P(D|M)*P(M)$$

This simply means that given a certain set of data, the *posterior* probability of a given model (i.e. its probability after the data has been observed) is proportional to the *prior* probability of the model (i.e. its probability before the observation of data) multiplied by the probability the model assigns the data. The P(D) term in Bayes' theorem that is eliminated from Bayes' rule can be thought of as normalising

---

<sup>21</sup> Probability is usually given an epistemic interpretation here – i.e. it expresses degree of belief (with 1 reflecting imperturbable certainty, and 0 unwavering disbelief), rather than, say, proportion of outcomes (as in the frequentist interpretation), or the apparently ontological stochasticity of quantum processes. This is necessary because the logic of Bayesian statistics seems to apply unproblematically to deterministic scenarios, as long as the observer is able to generate prior probabilities for different models, and knows what data each model predicts.

posterior probabilities such that they add up to one (note that this term remains the same regardless of the model under consideration).

## **Bayes Factors**

In practical terms, there are a number of ways of approaching Bayesian statistics, depending on the situation. If you have two well-defined alternative models  $M_1$  and  $M_2$  (one of which can be thought of as the null, if you wish), you can assign each a prior probability and then derive posterior probabilities using Bayes' theorem. However, it may be difficult to arrive at precisely quantified prior probabilities. If this is the case then Bayes factors can be used instead:

$$\text{Bayes Factor} = P(D|M_1) / P(D|M_2)$$

This does not correspond to the posterior probability of either of the models, but rather to the degree of support for one rather than the other. If we define prior-posterior probability ratios expressing the posterior versus prior probability for each model as follows:

$$PP_1 = P(M_1|D) / P(M_1)$$

$$PP_2 = P(M_2|D) / P(M_2)$$

Then it can be seen from rearrangement of Bayes' theorem that:

$$PP_1 / PP_2 = \text{Bayes Factor}$$

Bayes factors range from 0 to  $\infty$ . A Bayes factor  $> 1$  favours model 1, and a Bayes factor  $< 1$  favours model 2. Conventionally Bayes factors  $> 3$  or  $< 1/3$  are taken as evidence for one model over the other (Jeffreys, 1961; Kass & Raftery, 1995).

## **Bayesian Parameter Estimation**

Sometimes we may wish to perform Bayesian statistics without a clear idea of what the parameters for our alternative model should look like. In that case, we can perform Bayesian parameter estimation (Kruschke, 2011). The idea here is that rather than positing specific models with predefined parameter values, we simply define a prior distribution over a possible range of parameter values, and then derive a posterior distribution on the basis of the probability each parameter value assigns to the data we actually observe.

We can then reject or accept the null on the basis of the posterior distribution (Kruschke, 2011). If the posterior distribution's highest density interval (HDI – the highest average density continuous interval containing 95% of posterior probability distribution) lies close enough to the null value (i.e. within the region of practical equivalence, or ROPE), then we can accept the null. If all of the HDI lies outside the ROPE, then we can reject the null. If there is partial overlap, then more data may be needed to resolve the question.

As Bayesian parameter estimation is often most useful when we have no specific picture of what the alternative model ought to look like, we face the problem of choosing our prior distribution. The standard approach is to choose a weakly informative distribution – i.e. one that embodies some assumptions about the likely alternative models without being overly restrictive. If researchers want to allow the parameter to differ from the null in both directions, then they will choose a symmetrical weakly informative prior such as the Cauchy distribution, which resembles the normal distribution but with fatter tails and therefore somewhat more

weight on values distant from the centre (e.g. see Gelman, Jakulin, Pittau, & Su, 2008). The distribution can be centred on the null value, and given a scale that reflects the plausible range of values. For example, let's assume a binary predictor in logistic regression with values of -0.5 and +0.5. If the odds of one outcome vs. the other in the -0.5 condition are balanced, then a parameter value of 10 implies that in the +0.5 condition one outcome is over 20,000 times more likely than the other. If this magnitude of effect is unlikely, then the scale of the prior can be set appreciably lower than 10 (Gelman, Jakulin, Pittau, & Su, 2008 recommend 2.5).

### **Advantages of Bayesian Statistics over NHST**

Bayesian statistics have a number of advantages over NHST. Perhaps the most fundamental is that they directly address the question that researchers are (almost always) trying to answer: i.e. which model, theory, or hypothesis ought I to believe in given the results of this experiment? NHST, on the other hand, only directly tests the fit of data to a null, licensing rejection of the null when the fit is sufficiently bad. Sometimes this procedure does lead to normative reasoning (intuitively, if a model makes the wrong predictions, it's less likely to be true). However, it often falls short.

One reason is that NHST gives little sense of how strongly we should reassign plausibility from null to alternative hypothesis. Bayes factors on the other hand are a direct test of this. Another reason is that NHST ignores prior information: as far as NHST is concerned, if a  $p$ -value  $< .05$  establishes a psycholinguistic frequency effect, another  $p < .05$  in another experiment could equally well establish extra-sensory perception. Clearly though something has gone wrong here. One way of capturing what is to say that the prior on ESP is vastly lower than the prior on

frequency priming, an insight that Bayesian statistics can capture, but NHST cannot. A third weakness of NHST is that by failing to explicitly consider the predictions of alternative hypotheses, it may lead researchers to reject the null even in situations where the plausible alternative hypotheses explain the data just as badly, or worse.

However, the major advantage of Bayesian statistics for the purposes of this study is they allow us to establish support for the null hypothesis. In NHST, a  $p$ -value  $> .05$  could indicate that the data support the null, or could indicate that the study is currently underpowered. Bayes factors by contrast will tell you whether the data support the null, the alternative or neither (a result that can't be reliably achieved even combining NHST and power analysis – Dienes, 2014). Another advantage of Bayesian statistics for the current purposes is that though performing repeated inferential tests while adding more results to a dataset inflates the probability of type 1 error, Bayesian analyses avoid analogous problems (Kruschke, 2011). We will therefore submit Experiments 2-4 to a joint Bayesian analysis.

## **Bayesian Analyses**

As we do not have a clear idea of how big an effect of the interaction between condition and congruence to expect, and therefore cannot specify an alternative hypothesis in advance, we have opted for Bayesian parameter analysis (Kruschke, 2011 – see exposition above). In order to achieve this we use the R package *rstanarm* (Gabry & Goodrich, 2016). This package marries Bayesian statistics to the varieties of mixed effects model already available with *lme4* (Bates, Maechler,

Bolker, & Walker, 2015) that we have already been using, meaning that users can specify a model structure and a set of prior distributions for the parameters, and then estimate the posterior distribution on the parameters given a particular dataset. This is achieved via the earlier *rstan* package which provides an R interface to the C++ *Stan* library for Bayesian estimation (Stan Development Team, 2015), which itself uses Hamiltonian Monte Carlo to draw from the posterior parameter distribution, employing four chains per model.

**Priors** We based our priors on Gelman, Jakulin, Pittau, and Su's (2008) recommendations for weakly informative priors for Bayesian binomial models. All variables were centred at zero and scaled so as to have a standard deviation of 0.5. Priors (which were defined for the log odds ratios used as the models' parameters rather than for raw probabilities) took the form of Cauchy distributions. These are similar to normal distributions but have fatter tails, effectively assigning more prior plausibility to coefficient values further from the centre. All priors were centred at zero, corresponding to an assumption that an effect is equally likely in each direction (for predictors), or that the most likely level of performance is 50% correct (for the intercept). The scale (a measure of the spread of the distribution) was 10 for the intercept and 2.5 for the predictors, which effectively embodies the assumption that values greater than 20 (intercept) or 5 (predictors), or less than -20 (intercept) or -5 (predictors), are implausible.

Interactions are automatically assigned the same prior as individual predictors. This has the desirable side effect of making restrictions on interactions more stringent than those in main effects, and progressively more stringent the more main effects are involved in the interaction. For a binary predictor, which would be coded -0.5 vs.

0.5 in our scheme, the maximum plausible difference according to our prior would be:

*The difference between the coding levels \* the greatest plausible value of the coefficient*

i.e.

$$(0.5 - -0.5) * 5 = 5$$

However, for an interaction, the codes are simply the products of the codes for the predictors involved, therefore the code for a two way interaction varies between 0.25 ( $0.5^2$ ) and -0.25 ( $-0.5^2$ ), and thus the maximum plausible difference between interaction levels according to the prior would be:

$$(0.25 - -0.25) * 5 = 2.5$$

For a three-way interaction the maximum plausible difference would be halved again. Thus each additional effect involved in an interaction effectively halves the scale of the prior.

**Model Design** We opted to construct two models: Model 1 included all three experiments, whereas Model 2 excluded the last (which featured round names that may not have sufficiently contrasted with neutral names, as discussed above). Otherwise the models were identical. They were very similar to the models used for Experiments 2-4, except that a predictor was added for condition order (given that the bigger dataset allowed more predictors to be fitted), which is not of theoretical interest, but may clear up variance in such a way that differences between round and spiky conditions are easier to detect. A by-subjects random slope was also added for

condition order. All possible two- and three-way interactions were included (see Appendix 3.7 for full model specification).

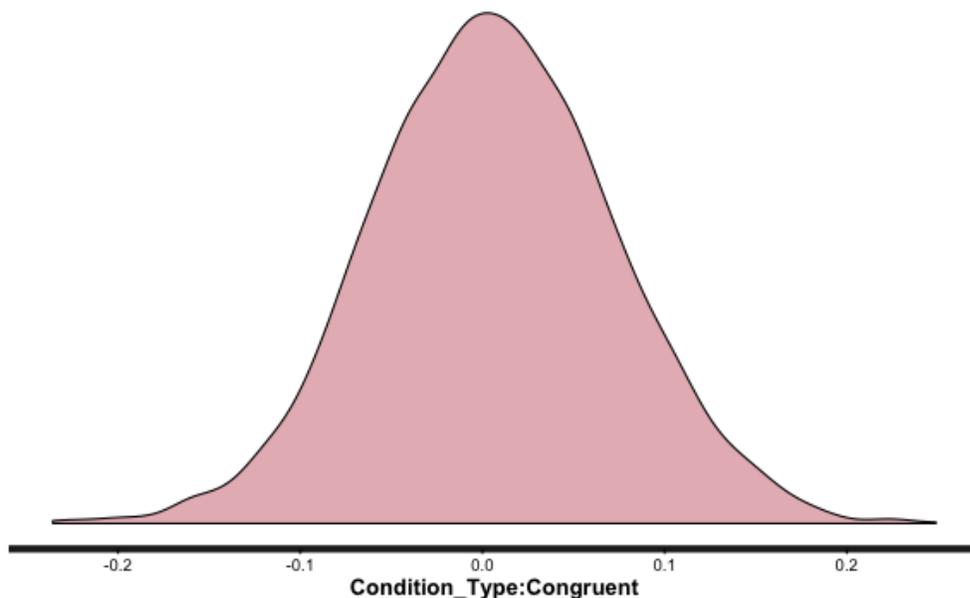
## Results

We will discuss results in terms of the 95% Highest Density Interval (HDI) of the posterior distributions produced by the model (see discussion above). Complete tables of posterior means and HDIs can be found in Appendix 3.8. Parameter estimates below also reflect the mean of the posterior distribution, and the distribution's HDIs. For each model I will discuss the predictors that the model suggests are reliable, and then examine whether condition type appears to modulate the effect of congruence and the interaction between congruence and category of foil.

**Model 1** - which encompasses Experiments 2-4, bears out the basic findings across the experiments presented in this chapter. For simplicity I will present as credible those results whose 95% HDIs exclude zero, as these results are simply sanity checks against the statistics presented earlier. There are credible effects of linear block ( $\beta = 1.498$ , 95% HDI [1.350, 1.653]) and quadratic block ( $\beta = -0.105$ , 95% HDI [-0.178, -0.030]), indicating that participants learned over the course of the experiments, and that this learning was faster between earlier blocks. There was a plausible effect of congruence ( $\beta = 0.123$ , 95% HDI [0.027, 0.216]), indicating that subjects performed better on congruent trials. There was a credible effect of condition order ( $\beta = 0.495$ , 95% HDI [0.360, 0.633]), indicating that participants tended to perform better on whichever condition came second (indicating a practice

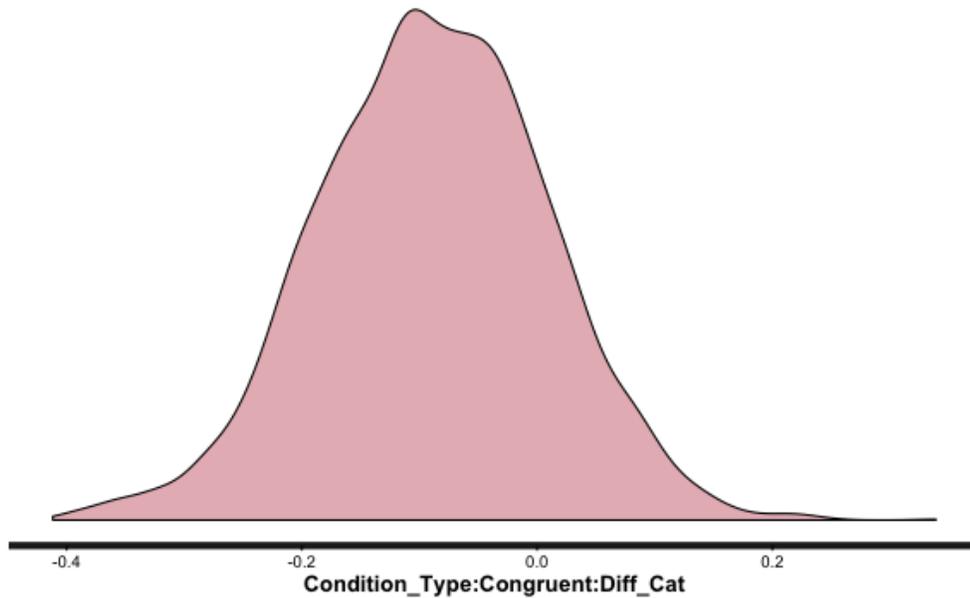
effect). As in some previous analyses there was a credible interaction between congruence and category of foil ( $\beta = 0.326$ , 95% HDI [0.220, 0.431]), indicating that the advantage of congruence was more pronounced for trials where the foil was from the opposite category to the target. There were also credible interactions between condition order and both linear ( $\beta = 0.325$ , 95% HDI [0.192, 0.4520]) and quadratic ( $\beta = -0.202$ , 95% HDI [-0.316, -0.091]) block, indicating that learning happened faster in whichever condition came second (again indicating a practice effect).

There are also difficult-to-interpret three-way interactions that appear (narrowly) credible: an interaction between quadratic block, condition order, and congruence ( $\beta = -0.222$ , 95% HDI [-0.436, -0.007]), and an interaction between condition order, congruence, and category of foil ( $\beta = -0.271$ , 95% HDI [-0.476, -0.059]). This latter may indicate that congruence (the advantage of which is mainly manifest when the foil and target are from different categories) is of more advantage in whichever condition comes first.



*Figure 3.15: The estimated posterior probability distribution for the interaction between congruence and condition type in Model 1. The x-axis represents parameter values, the y-axis probability density*

Now we turn to the crucial interaction between congruence and condition type. The interaction was centred very close to zero, with narrow HDIs compared to the prior ( $\beta = 0.006$ , 95% HDI [-0.122, 0.139]) – see Figure 3.15. To give a sense whether this interaction is likely to meaningfully modulate the overall effect of congruence, we can assume that the most extreme value in the HDI is correct and ask whether even then the interaction would make an appreciable difference. If we take the intercept as our baseline (as predictors are centred on zero the intercept reflects the typical predicted accuracy across the whole experiment) and assume that the correct parameter value for the interaction is 0.139, then the predicted average accuracy for the higher coding level for the interaction would be 82.5%, whereas predicted accuracy for the lower coding level would be 81.5% - i.e. the predicted difference is one percentage point. This is in contrast to a rather bigger main effect of congruence: assuming the mean of the posterior, overall accuracy for congruent trials is 82.9%, whereas the overall accuracy for incongruent trials is 81.1%.

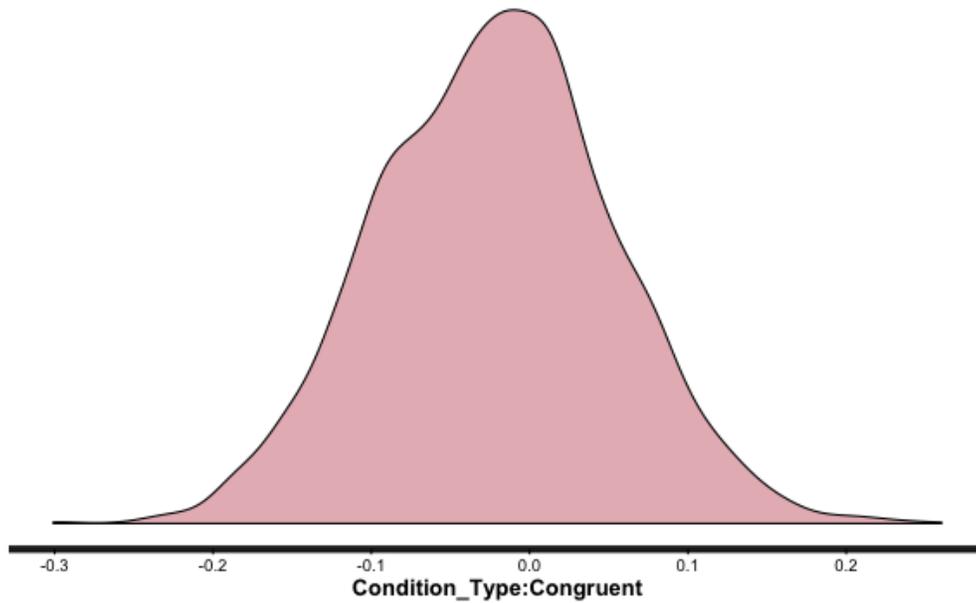


*Figure 3.16: The estimated posterior probability distribution for the interaction between congruence, category of foil, and condition type in Model 1. The x-axis represents parameter values, the y-axis probability density*

Similarly, the posterior HDIs for the interaction between congruence, category of foil, and condition type include zero, and are narrow compared to the prior ( $\beta = -0.089$ , 95% HDI [-0.284, 0.098]) – see Figure 3.16. Unlike with the previous interaction, the mean of the posterior is somewhat lower than zero, squaring with the result of Experiment 2 (i.e. that congruence – here modulated by category of foil – is a bigger advantage in the round condition). Again assuming the largest absolute value contained in the HDIs (-0.284) is correct, we can look at overall performance in each of the coding levels of the interaction. Performance at the higher level is predicted to be 82.5%, whereas performance at the lower would be 81.5%. By contrast, the mean of the posterior for the interaction between congruence and category of foil predicts that typical performance at its higher level would be 83.2%, whereas performance at its lower level would be 80.8%.

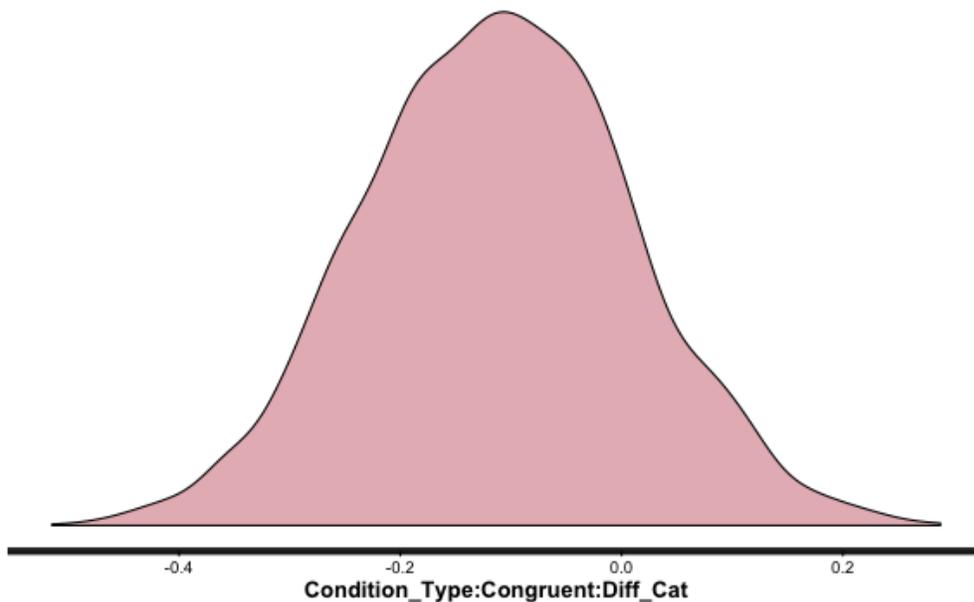
It is debatable whether the extremes of the HDIs for these two interactions of congruence with condition type fall inside the region of practical equivalence with zero. On the one hand they reflect smaller effects of congruence than apply across both conditions. But on the other they are within the same order of magnitude. It may be the case that we do not have enough data to definitively rule out the existence of these interactions. However, they are clearly less credible than the cross-condition effects, and our results are certainly consistent with their non-existence.

**Model 2** - which excludes Experiment 4, largely reflects the results of Model 1. Once again there are credible effects of linear ( $\beta = 1.638$ , 95% HDI [1.45, 1.835]) and quadratic ( $\beta = -0.155$ , 95% HDI [-0.255, -0.06]) block, condition order ( $\beta = 0.472$ , 95% HDI [0.322, 0.620]), and category of foil ( $\beta = 0.091$ , 95% HDI [0.019, 0.164]) – indicating that participants performed better on trials where the target and foil shapes came from different categories. This time, the HDIs for the main effect of congruence do encompass zero ( $\beta = 0.087$ , 95% HDI [-0.022, 0.198]), albeit narrowly. However, as before, there is a credible interaction between congruence and category of foil ( $\beta = 0.242$ , 95% HDI [0.114, 0.369]), indicating that an advantage for congruence is present when the target and foil are from different categories. Once again there were interactions between condition order and both linear ( $\beta = 0.329$ , 95% HDI [0.196, 0.462]) and quadratic ( $\beta = -0.213$ , 95% HDI [-0.333, -0.094]) block. As with Model 1, there was a difficult-to-interpret interaction between quadratic block, condition order, and congruence ( $\beta = -0.253$ , 95% HDI [-0.498, -0.009]).



*Figure 3.17: The estimated posterior probability distribution for the interaction between congruence and condition type in Model 2. The x-axis represents parameter values, the y-axis probability density*

Turning to the crucial interactions of congruence with condition type: again, the posterior distribution for the interaction between congruence and condition type is narrow compared to the prior, and centred close to zero ( $\beta = -0.022$ , 95% HDI [-0.165, 0.125]) – see Figure 3.17. Following the same procedure as for Model 1 we see that if we assume the largest absolute value in the HDI is correct, and take the intercept as our baseline, the difference between the levels of the interaction is 84.2% versus 85.3%. This is the same as the difference when the main effect of congruence is examined in the same way, assuming the mean of the posterior.



*Figure 3.18: The estimated posterior probability distribution for the interaction between congruence, category of foil, and condition type in Model 2. The x-axis represents parameter values, the y-axis probability density*

The three-way interaction between congruence, category of foil, and condition type is also narrow compared to the prior, and again encompasses zero ( $\beta = -0.112$ , 95% HDI [-0.350, 0.125]) – see Figure 3.18. As in Model 1, the mean is below zero, reflecting Experiment 2’s finding that congruence is a bigger advantage in the round condition. Assuming the highest absolute HDI, predicted average performance at the two levels of the interaction varies between 84.2 and 85.3%. Assuming the mean of the posterior distribution for the two-way interaction between congruence and category of foil, we predict average performance of 84.0% at one level of the interaction as against 85.5% at the other.

Compared to Model 1, Model 2’s 95<sup>th</sup> percentiles for the HDIs of the interactions involving condition type and congruence were bigger relative to the effects of congruence across blocks. This presumably reflects the fact that excluding

Experiment 4 from Model 2 gives more weight to Experiment 2, the only individual experiment with a marginal block-congruence interaction. However, once again this is only the case if we assume that the extreme values of the HDIs are correct. Thus there is clearly less support for the interaction between congruence and condition type than for the effect of congruence across conditions, and our results are consistent with their being no congruence-condition interaction.

## Discussion

Here I have presented evidence across four experiments that iconicity enhances performance in a statistical learning paradigm. Close analysis of the beginning of Experiment 1, and the consistent tendency for iconicity to be a greater advantage when the foil presented during the trial does not also match the name, suggest that the benefit of iconicity in these experiments is to do with picking out the right referent during a particular trial, rather than allowing faster or more robust memory encoding of name-shape pairings across trials. If iconicity helps bring referents to the mind's eye, then it is unsurprising that it should confer this benefit of disambiguating reference.

However, the *bona fide* learning here is a result of cross-situational statistics, and if iconicity does not facilitate this learning (as the lack of interactions between iconic congruence and rate of learning seems to suggest), then how important can we truly claim iconicity to be in this kind of process? It seems only to be a back-up cue used when participants would otherwise have to guess.

In the current set of experiments only two referents are potential matches for a name in a given trial. Therefore an ideal learner could learn any name after no more than two exposures (assuming the two trials have different foils). This means that learning on the basis of cross-situational statistics is quick and easy, and that iconicity (which, as participants may also learn, is not actually informative as to shape identity) is redundant. My prediction is that iconicity would start to become much more crucial to learning if we made the problem of referential ambiguity harder, and if we allowed most iconicity to be congruent (as it is in natural language, rather than equally balanced between congruence and incongruence - i.e. anti-iconicity - as is the case here), making reliance on iconicity a consistently fruitful strategy. In crowded visual scenes where brute-force cross-situational learning may be hampered by the memory demands posed by the number of potential referents, iconicity could be highly effective in reducing referential ambiguity. Moreover, iconicity may be even more valuable in non-ostensive learning contexts, i.e. those where the referent (be it an object, action, or event) is entirely absent. If the referent of the word is not actually present, then cross-situational learning of the kind seen here ceases to be possible. However, if iconic speech, gesture, or sign can evoke properties of the absent referent, then learning can take place in spite of its absence.

The second set of findings from the experiments presented in this chapter relate to the relative importance of rounded-rounded and spiky-spiky mappings in sound-shape iconicity. Some results presented in other chapters hint that the rounded association might be stronger than the spiky one (see Chapter 4). This is interesting because if it were true, it might help elucidate the mechanism of sound-shape iconicity. For instance, one hypothesis is that rounded sounds are associated with rounded lip shape, and that the iconicity arises from cross-modal sound-shape

correspondences during speech production and comprehension (Ramachandran and Hubbard, 2001). If this is indeed part of the mechanism for sound-shape iconicity, we would expect rounded associations to be primary, and spiky associations to arise later through something like a principle of contrast. If this were indeed the case then rounded associations should plausibly be stronger than spiky associations.

Three of the experiments in this chapter (Experiments 2-4) are designed to test this possibility, by sequestering round and spiky iconicity into two separate conditions, and seeing whether iconic congruence exerted a stronger effect in one or the other. Experiment 2 appeared to suggest (with marginal significance) that iconicity improved performance in the round condition, but not in the spiky condition. However, this asymmetry failed to replicate in the remaining two experiments. In order to test whether there really is an asymmetry, we submitted data from all three experiments to a Bayesian analysis. This both obviated problems with pooling and reanalysing data using classical inferential statistics, and allowed us to demonstrate positive support for the null hypothesis, if such support was present.

Though the results from the Bayesian analysis were somewhat inconclusive, they suggest that any asymmetry between the conditions is a smaller effect than the overall influence of iconic congruence across conditions, and indeed they are consistent with there being no asymmetry at all. How to reconcile this with results from other chapters is unclear: leaving aside the possibility of Type I or Type II error, it could be the case that there is a round-spiky asymmetry that is stronger in production than in comprehension, because spoken production inevitably involves lip rounding during the appropriate phonemes, whereas listening to speech may activate much weaker representations of lip shape. Indeed, it would be interesting to

test whether a round-spiky asymmetry *is* observed in cross-situational learning if participants are able to see the mouth of the person producing the names, or are encouraged to mouth along themselves. However, it would seem that while lip shape might be part of the story for the mechanism of sound-shape iconicity, it does not represent a complete explanation.

# Chapter 4: Iconicity in Language Evolution<sup>22</sup>

## Introduction

This chapter presents results from a series of experiments focused on the spontaneous emergence of iconicity in word invention and production. The mainstay of the chapter is two experiments that employ the iterated learning methodology (Kirby, 1999; Kirby, 2000; Kirby, 2002; Kirby, Cornish, & Smith, 2008; Kirby & Hurford, 2002; Kirby, Smith, & Brighton, 2004; Thompson, Kirby, & Smith, 2016), a model of cultural evolution. These are followed by a simple word creation experiment to help clarify the iterated learning results. To lay the groundwork for these experiments I will first discuss iterated learning and its theoretical motivation.

To preview what I argue below, one crucial theoretical problem in the study of iconicity is how it ends up in vocabularies at all. As emphasised in Chapter 1, the mere fact that it seems to be useful to processing and acquisition is not a complete explanation of this. I argue that this problem can be solved by linking iconicity's presence in lexica with another little-studied facet of iconicity: its role in language change. My claim (following previous arguments from Kirby et al. about key aspects of grammar) is that iterated learning explains the presence of iconicity in the lexicon: defeasible acquisition and production biases are amplified across a process of cultural transmission, creating a pressure for lexica to incorporate and maintain iconicity.

---

<sup>22</sup> Thanks to Alex Lau-Zhu for constructing the LetterScore norm, Alex Lau-Zhu and Gabriella Vigliocco for input to the concept for Experiment 1 and 2, Julio Santiago and David Vinson for assistance with experimental design, David Vinson for statistical advice, and Nourane Clostre for help with data collection for Experiment 1.

In order to clarify the claim I am making and substantiate its plausibility, I will begin with a fairly extensive introduction to the literature on iterated learning, explicating its logic and showing what it has already achieved in the study of grammar (as opposed to vocabulary).

## **Iterated Learning**

Iterated learning is a paradigm whereby agents (either adult: Kirby, Cornish, & Smith, 2008; Smith & Wonnacott, 2010; Verhoef, Kirby, & de Boer, 2014; child: Kempe, Gauvrit, & Forsyth, 2015; animal: Claidière, Smith, Kirby, & Fagot, 2014; or artificial: Brighton & Kirby, 2006; Brighton, Smith, & Kirby, 2005; Kirby, 2002; Kirby, Dowman, & Griffiths, 2007) learn from other agents who themselves learned in the same way (Scott-Phillips & Kirby, 2010). Pioneered by Simon Kirby and collaborators (Kirby, 1999; though it was independently prefigured in some respects by Esper, 1966<sup>23</sup>), it comes in several forms depending on the aspect of cultural transmission that researchers are interested in (see Kirby, Griffiths, & Smith, 2014, for a recent review). Broadly speaking, horizontal versions of the paradigm involve information exchange relationships within dyads or larger groups of agents, whereby within each interaction, each agent is both a learner and a teacher (e.g. Galantucci & Garrod, 2011; Garrod et al., 2007; Scott-Phillips, Kirby, & Ritchie, 2009; Theisen et al., 2010). Vertical forms by contrast feature relationships where transmission is all one way: e.g. participant 1 teaches participant 2, who teaches participant 3, who teaches participant 4, etc. (e.g. Kirby, Cornish, & Smith, 2008; Smith & Wonnacott, 2010; Verhoef, Kirby, & De Boer, 2014). It is vertical forms that have been used most

---

<sup>23</sup> Credit for my being aware of this reference goes to Morten Christiansen's remarks at the Fifth Implicit Learning Seminar: June 2016, Lancaster, UK.

extensively, and by default this will be the design I mean when I refer to iterated learning (though some studies have combined both horizontal transmission within generations and vertical transmission between them for maximum ecological validity: Theisen-White, Kirby, & Oberlander, 2011).

In principle iterated learning can be used to study any culturally transmitted information or behaviour (e.g. transmission of geometrical designs: Claidière, Smith, Kirby, & Fagot, 2014; Kempe, Gauvrit, & Forsyth, 2015). However, language is the topic it was originally designed to study, and the topic it has been most frequently applied to since.

### **Biological Evolution and the Cultural Evolution of Language: Compare and Contrast**

The study of language change is not new. The long tradition of historical linguistics deals with language change over time: as early as the seventeenth century it was speculated that Latin, Greek, and Hindi shared a common linguistic ancestor (Campbell & Poser, 2008). Much has been learned about language change from a historical point of view, and it is well established that change is not entirely arbitrary (Campbell, 1998).

The novel idea embodied by the iterated learning approach is that languages do not merely evolve, and do not merely evolve in (somewhat) predictable ways, but that languages evolve under pressure to become *fitter* in something like the Darwinian sense. Moreover, the claim is that much of the structure within language and many of the commonalities between languages are due to this factor (Christiansen &

Chater, 2008; Kirby, Smith, & Brighton, 2004). The analogy between language evolution and biological evolution goes back at least as far as Darwin (1871)<sup>24</sup>. However, the current interest mirrors a recent turn towards cultural evolutionary modes of explanation in the study of culture and behaviour (e.g. Barrett et al., 2016; Blackmore, 1999; Richerson & Boyd, 2004; Rogers & Ehrlich, 2008).

What does it mean for a culturally transmitted system of knowledge like a language to be evolutionarily fit? The best way of explaining this is to look at the definition of natural selection that biologists work with (Darwin, 1859), and then to explain its similarities and differences with the principles at work in language evolution.

A population of replicators (e.g. bacteria, plants, animals) undergoes natural selection when the following conditions are met:

- 1) Variation: members of the population must vary on some trait.
- 2) Heredity: offspring must resemble their parent(s), effectively inheriting their variant(s) of traits.
- 3) Differential fitness: some members of the population must enjoy greater fitness – i.e. reproductive success - than others as a result of variation in heritable traits.

If these conditions are met, then organisms with variants of heritable traits that give the highest fitness will out-reproduce other organisms. They will pass traits onto their offspring (including the trait responsible for the superior fitness), who will constitute a greater proportion of their population than their parents did. In the simplest scenario, this process is repeated generation on generation until the entire population is made up of individuals with the fittest variant of the trait.

---

<sup>24</sup> Indeed Gontier (2011) suggests that tree diagrams illustrating historical relations between languages inspired Darwin's own phylogenetic tree diagrams in the *Origin of Species*.

To take a real-life example of classic natural selection from Ricker (1981): in the 1950s fisherman in British Columbia started to catch pink salmon using gill netting, which traps bigger fish but lets smaller ones through. There was variation in the size of the salmon in the population, and much of this variation was heritable, i.e. a result of the salmon's genes. After the netting was introduced, there was an immediate fitness cost to being big. By the same token the relative fitness of small salmon increased. Small salmon outcompeted big ones in survival and therefore reproduction, passing on the heritable part of their smallness. The result was that over the course of 25 years the average weight of a salmon decreased by one third.

What then would selection mean as applied to culture? In this instance, the replicator is argued to be some (at least partly) culturally determined convention or practice. This could be a way of making canoes (Rogers & Ehrlich, 2008) or a stance towards iniquity in economic transactions (Barrett et al., 2016), but for our purposes we can say that this replicator is a word in the mental lexicon. Its environment is the mind of the people who know it. It replicates when someone new learns it on the basis of language input from somebody who already knows it.

Does this replication meet the criteria for selection? Clearly it involves substantial heritability: if the new learner acquires the word successfully (as learners typically will), then the entry in her mental lexicon will have inherited its properties from its 'mother' in the first person's lexicon. What about variation and variable fitness? Suppose we define the population for this particular word as the set of entries in people's lexica that share the same meaning, while potentially differing syntactically, morphologically and phonologically. To take a simple real life example, let's look at the population of the word *formula*. There are two variants in this population: Variant 1, where the plural is *formulas*, and Variant 2, where the plural is *formulae* (following

the Latin). In the first case the plural is simply the result of the application of the normal English plural morphology, whereas in the second case the plural form has to be stored partially suppletively. It is this property that introduces the difference in fitness. Variant 1 is acquired easily when heard, whereas Variant 2 relies storing extra information in memory, which may not happen successfully unless the learner has regular exposure to Variant 2's plural form. And indeed, if we look at the relative frequency of the two plural forms in historical text corpora, we find that in 1900 they are at roughly equal frequency, whereas by 2000 Variant 1 is more than twice as frequent as Variant 2 (source: Google books ngram viewer: Michel et al., 2011).

At this point there are at least three qualifications that need to be made regarding the analogy between biological natural selection and cultural transmission of language. The first is that there is no analogue for the genotype-phenotype distinction (this point is also noted by Brighton, Smith, & Kirby, 2005). If we treat cognitive representations as the equivalent of organisms, then there is nothing we can treat as the equivalent of genes. However, though this is a non-trivial distinction with the mechanism of biological natural selection, it does not undermine the claim that Darwinian processes are at work. Natural selection does not depend on a genotype-phenotype distinction (no mention of one is made in the conditions for natural selection outlined above). When Darwin (1859) first introduced the idea, he was not aware of genetics, and in fact in later editions of the *Origin* Darwin even flirted with Lamarckian ideas by suggesting that acquired characteristics might “top up” variation within species (Larson, 2004). None of these mechanisms of inheritance are inconsistent with the claim that natural selection is in operation.

Another point of difference is that whereas selection on organisms tends to operate on both survival and replication (i.e. reproduction), in language evolution its role is

largely limited to replication (again, Brighton, Smith, & Kirby, 2005, note this point). Some linguistic forms and structures may be easier to acquire than others (i.e. be more adaptive for replication), but once learned, survival is not generally a problem (assuming that learners do not subsequently forget what they have learned). Again, this is a non-trivial distinction to be borne in mind, but remains consistent with the definition of natural selection.

A more crucial point is that much of the change we are interested in might look more like *selective mutation* than selection proper, which has to involve competition (cf. Brighton, Smith, & Kirby, 2005, on reanalysis; and Claidière, Kirby, & Sperber, 2012, on the importance of directed mutation in cultural evolution, alongside selection). For instance, suppose a language learner only ever hears the word *formula* from their maths teacher in its singular form. The teacher's lexicon may contain Variant 2 of the word (plural: *formulae*), but barring something exceptional, the learner will assume that the plural of the word is *formulas*, as normal English morphology dictates, and will therefore acquire Variant 1. The entry in the teacher's lexicon is clearly the cause of the entry in the learner's lexicon, and therefore the learner's entry is the offspring of the teacher's entry in our analogy. Thus in course of replicating itself, Variant 2 has mutated into Variant 1. For obvious reasons, it is highly unlikely that this process would happen the other way round. Therefore this process can be expected to result in systematic language change without involving natural selection proper.

This process is unlike what is seen in biological evolution, where mutation is typically random (albeit in quite a subtle relationship with selection: bacteria can selectively regulate their rate of selection to optimise it for different conditions – Galhardo, Hastings, & Rosenberg, 2007; and transcription factor genes like the Homeoboxes can lead to major changes in body plans resulting from small numbers of mutations).

Nonetheless, this kind of directional mutation looks quite similar to selection for our purposes: both will drive aspects of the language into certain non-arbitrary parts of the overall space of possibilities. In iterated learning, it may not always be obvious whether competition or selective mutation is responsible for the evolution we see. Indeed, if you treat the *whole* language as the unit of selection (as many do: e.g. Brighton, Kirby, & Smith, 2005), then the debate only makes sense if it is framed in terms of mutation rates – fit languages are not those that outcompete rivals (there are no rivals), but rather those that mutate slowly (i.e. remain stable), even with a transmission bottleneck. Brighton, Kirby, & Smith (2005) note that “in this context [of cultural selection for learnability], the terms adaptive and selection only loosely relate to the equivalent terms used in the theory of biological evolution.” In the discussion we will return to the question of whether competition or mutation is driving the patterns seen in the experiments I present in this chapter.

### **Cultural Evolution, and Language Universals, Acquisition, and Structure**

Why take cultural evolutionary view of language? The claim is that rather than simply explaining neat but relatively trivial facts about vocabulary change (like the *formula* example above), cultural evolution plays a crucial role in explaining language’s most fundamental properties (Beckner et al., 2009; Kirby, Smith, & Brighton, 2004; Christiansen & Chater, 2008). This stance needs to be understood in the context of the previous sixty years of debate over innateness, language acquisition, and language universals.

It is uncontroversially the case that for a large number of surface properties of language (e.g. word or phrase order), the distribution of actually observed

configurations in the world's language is highly skewed towards some logical possibilities over others, with some absent entirely (Greenberg, 1963)<sup>25</sup>. Moreover, some traditions, particularly those influenced by generativist nativism, claim that at a properly abstract level of analysis, there are strict universals regarding phenomena like constituent categories, relationships between constituents in hierarchical structure, and syntactic movement (see e.g. Pinker, 1994; Pinker & Bloom, 1990; though note that others vehemently contest the existence of such universals: Evans & Levinson, 2009). Following Chomsky, the typical generativist account of these universals appeals to Universal Grammar (UG), innate knowledge of the structure of language (see Chomsky 1972; 1981, for classic formulations)<sup>26</sup>. Thus the commonalities between the superficially disparate languages of the world are the result of the contents of UG, and indeed are so strong that "According to Chomsky, a visiting Martian scientist would surely conclude that aside from their mutually unintelligible vocabularies, Earthlings speak a single language" (Pinker, 1994, p. 232). Thus both language acquisition and language universals are explained on the basis that children have innate knowledge of language, and could not possibly learn

---

<sup>25</sup> Greenberg's universals are largely of the form: *If a language possesses property A, it will also possess property B*. For flavour, here is Universal 20: "When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or its exact opposite."

<sup>26</sup> The UG terminology is used inconsistently, and depending on the text may refer to common properties of all languages; "state-zero" of the language acquirer's brain; or state-zero of the brain plus a "language acquisition device" that moves from state-zero to the final target grammar (Kirby, Smith, & Brighton, 2005; Jackendoff, 2002). The latter two interpretations psychologise UG, and if one of them is adopted, there are additional decisions to be made over whether this knowledge of language is taken to be explicit and propositional, or implicit in the representational capabilities or acquisitional abilities of the language system. However, on any interpretation, UG's upshot is that: (1) human beings are only capable of acquiring language by virtue of UG, (2) human beings can only acquire languages of the class defined by a certain set of properties (those encoded in UG), and (3) therefore everyone with linguistic competence will represent their language as having these properties.

a language that does not conform to that innate knowledge while easily acquiring a language that does<sup>27</sup>.

Proponents of the cultural evolutionary view take a different stance. They suggest that rather than representing immutable, hardwired constraints on the form language can take, language universals are in fact the result of acquisition biases amplified by cultural transmission (Kirby, Smith, & Brighton, 2004; Culbertson & Kirby, 2016). Instead of the brain evolving to shape itself to language via the slow process of biological evolution, language may evolve to shape itself to the brain through the much faster process of cultural evolution. Indeed modelling suggests it would be extremely difficult for natural selection to fix linguistic conventions in the genotype, because culture evolves much more quickly than organisms (Christiansen & Chater, 2008).

Rather than the “hard” constraints of classic UG – principles that admit of no exceptions – the biases that shape language may be “soft”, i.e. relatively weak biases or learnability advantages for one linguistic possibility over another (Thompson, Kirby, & Smith, 2016). This means that humans may be capable of acquiring a far greater range of languages than are actually observed, because if outlandish languages ever arose, even though they would be learnable (with difficulty), acquisition biases would quickly drive the cultural evolution of language into a more learnable subset of the language space<sup>28</sup>. Moreover these soft biases

---

<sup>27</sup> It should be pointed out that opinions differ among proponents of UG as to whether it represents an adaptation for establishing communication (Pinker & Jackendoff, 2005), or whether its origins are to be explained in terms of something other than natural selection (Chomsky, 1988).

<sup>28</sup> It is not necessarily the case that UG and cultural evolution are mutually exclusive explanations for language structure and universals. Depending on one’s construal of UG, it could be the case that UG defines a space of possible languages, and cultural evolution systematically directs languages into certain parts of that space. It could also be the case that what generativists speak of as UG in fact constitutes some of the biases shaping cultural

may be domain-general rather than domain specific (Culbertson & Kirby, 2016; Christiansen & Chater, 2008). Though generativist discussions of language have traditionally denied this, and supported claims for autonomous linguistic principles by focusing on intricate and apparently radically arbitrary language phenomena such as so-called wh-islands (Ross, 1967) - which seem difficult to explain on functionalist or emergentist grounds - there are signs that in recent years even the official generativist position has become more amenable to the idea that UG is relatively minimal, and supplemented by domain-general considerations (Chomsky, 2005; Hauser, Chomsky, & Fitch, 2002).

Crucial to the expression of these biases is the transmission bottleneck (Hurford, 2002; Kirby, Smith, & Brighton, 2004). When a learner is acquiring a language, they do not receive the whole language in their input, but rather only a subset. This subset underdetermines the language (i.e. set of possible well-formed expressions), and moreover the language underdetermines the grammar (i.e. set of rules) that produces it (as multiple grammars may produce the same language). Given a

---

evolution (both possibilities are considered in Brighton, Kirby, & Smith, 2005, though deemed problematic for proponents of UG; cf. Kirby & Hurford, 1997, for the latter).

Alternatively UG could be seen as the representational capabilities used for mental grammars, which will in turn constrain what grammars can and will be learned. E.g. Kirby (2002) has artificial agents represent simple languages as context free grammars (CFGs), and shows that compositionality and recursion emerge in the course of iterated learning. Kirby points out that features like compositionality and recursion are not automatically built in to grammars of these forms. Nonetheless, once the agent can represent CFGs then the power to represent compositionality and recursion has already been granted, and in that sense is unlearned. However, the key point Kirby is able to make is that agents converge on recursively compositional grammars on the basis of a combination of the right representational format and quite general description length principles, without any built-in domain specific knowledge about what the grammar should look like.

So to return to the question of the relationship of UG to cultural evolution, let us suppose that UG is simply construed as the ability to represent a grammar that is compositional and recursive, perhaps with features like constituency, headedness, and infinite generativity. Some aspect of this ability must at bottom be innate on any cognitivist view: see Fodor (1980), Tenenbaum, Kemp, Griffiths, & Goodman (2011). Once it is granted that mental grammars do have these properties (as many proponents of cultural evolution would happily do), then it seems the debate over UG ceases to be about much more than terminology.

(plausible) bias on the part of the learner for simple grammars that describe their language in terms of a relatively small number of lexical items and rules rather than a large number of arbitrary idioms, some grammars are potentially learnable from this subset, whereas others are not. Specifically, a large (indeed infinite) language can be learned on the basis of a comparatively small subset of expressions if the language can be described via a compositional grammar: if every expression is a function of the meaning of its components and their combination, then rather than learning every expression by rote, the learner can learn lexical meanings and compositional rules on the basis of a subset of strings, and then generalise to unseen strings. However this is impossible with a language of syntax-free idioms. Thus idiomatic languages are inherently unstable, whereas compositional languages (or even compositional principles within a largely idiomatic languages) are not, and therefore cultural evolution will tend to fix the latter, moving the language towards compositionality (Kirby, Smith, and Brighton, 2004).

### **Evidence for the Power of Iterated Learning**

The iterated learning thesis has been fleshed out across a number of studies. Experiments with artificial agents demonstrate that in principle iterated learning can result in structured language, without that structure being built in to agents' prior (i.e. innate) knowledge. Kirby (2002) set up a diffusion chain of computer agents each faced with the task of learning a grammar based on incomplete input of form-meaning mappings, and a requirement to capture that input with as simple a grammar as possible. Later generations evolved grammars that were not only compositional but recursive. Brighton, Smith, & Kirby (2005) found that when using

computer agents with a tendency to generalize based on minimum description length principles for representations of form-meaning mappings, compositionality emerged spontaneously through iterated learning. Brighton & Kirby (2006) found that diffusion chains of computer agents converged on globally coherent signal-meaning mappings. Kirby, Dowman, & Griffiths (2007) found that iterated learning amplified even a tiny bias towards systematicity on the part of computer agents into a completely systematic set of form meaning mappings. Thompson, Kirby, & Smith (2016) show that weak biases towards certain syntactic parameter settings over others can result in behavioural syntactic universals, and that such universals are unlikely to then be fixed in the genes in an instance of the “Baldwin Effect” (genotypic fixing of phenotypic properties that were previously induced by the environment).

In each of these experiments, while communicative efficiency may have increased, the driver for the emergence of structure was an increase in learnability: i.e. linguistic fitness. Given an entirely biologically plausible preference for regularity and parsimony in grammar, the computer agents in early generations picked up on chance regularities in the incomplete input they had access to, built them into their ‘grammar’, and produced minor regularities in their own output, which were then amplified into total regularity by subsequent generations. The information bottleneck represented by incomplete input data (analogous to the putative poverty of the stimulus presented as evidence for innate generative structures) seems to be not just consistent with this process but a prerequisite for it (Cornish, Tamariz, & Kirby, 2010).

More recently, a number of iterated learning experiments have been performed using human participants. Kirby, Cornish, and Smith (2008) demonstrated that when you place human participants in a vertical iterated learning experiment, an initial arbitrary

language (i.e. one without iconicity or compositionality) will develop morphology-like properties. This is not a result of conscious design on the part of the participants (who report being unaware of introducing it), but is rather the result of nascent morphology appearing by accident (due to participant's mistakes), and being retained because it is easier to remember than arbitrary names. This evolution is not driven by language specific considerations, but rather by a very general feature of cognition: limited memory capacity. Smith & Wonnacott (2010) found that diffusion chains but not isolated learners regularised plural markers in a semi-artificial miniature language. Verhoef, Kirby, & de Boer (2014) conducted a study where participants learned a language made by using a slide whistle – a simple and intuitive musical instrument. This had the advantage of keeping signal production easy while making it less likely that performance could simply piggyback on pre-existing competence in natural language. Each participant had to learn and reproduce twelve whistles. As with the previous syllable based experiment, structure spontaneously emerged. By the end of the diffusion chains, whistles had evolved into constructions involving a limited number of sonic building blocks – essentially whistle phonemes. This corresponded to an increase in accuracy of recall over the course of the chains. Whistles that could be coded as a combination of standardized short sounds were easier to remember than whistles that had to be recalled holistically – thus yet again the evolution seen can be thought of in terms of fitness of language.

Thus to summarise, iterated learning represents an alternative approach to linguistic nativism for understanding language structure and universals. There now exists a

wide range of theoretical and empirical evidence that it is capable of explaining important features of language.

So far its proponents' focus has tended to be on grammar, particularly syntax, as for the last 60 years this has been the arena where the fiercest battles over the nature of language have been fought. However, there is no principled reason why the iterated learning approach can't yield insights into the vocabulary too. It is true that vocabulary presents a different kind of problem for the learner: much of the action in iterated learning models of syntax change relates to the transmission bottleneck: the fact that the learner cannot be exposed to every expression in the language, and therefore has to derive a grammar from an incomplete data set. With vocabulary the situation is different – much more extensive coverage can be achieved in the input. However, input isn't perfect, and clearly natural language vocabulary does evolve. General principles of iterated learning (cultural evolution proceeds according to systematic biases as well as random drift) are applicable. If certain types of word are more learnable, or if learning and production mistakes consistently throw up certain kinds of words, we would expect language evolution to experience a pressure towards these kinds of words.

if - as Thompson, Kirby, and Smith (2016) argue - many language universals are the result of soft, domain general biases amplified by cultural evolution, and if human beings are biased towards iconicity in learning and production (as established in Chapters 1 and 2), this opens up the possibility that iconicity can be considered a universal on the same terms. In support of this, we can note that though iconic wordforms may be in minority in the world's lexica (Newmeyer, 1993) they seem to be ubiquitous. Moreover languages seem to incorporate iconicity to the greatest

extent they can: sign languages, with their much richer iconic potential, incorporate abundant iconicity.

The suggestion is not of course that all words are iconic, or that all languages are constantly becoming more iconic: iconicity is merely one of several pressures at work, and will eventually reach a state of equilibrium. Perniss and Vigliocco (2014) have argued that iconicity and arbitrariness coexist in language as adaptations to different basic pressures: words need to effectively activate representations of their referents for communication, but different words must also be discriminable in the speech stream. While iconicity evokes absent referents (establishing displacement in evolution and development); arbitrariness aids discriminability, and thus learning and processing efficiency (Hockett, 1960; Monaghan, Christiansen, & Fitneva, 2011).

However, our prediction is that where possible, iconicity will tend to accrue in the lexicon through a process of cultural evolution, on the basis of domain general biases. I explore this possibility in the following experiments. Experiments 1 and 2 show that these biases result in the emergence of iconicity in a model of language change, with Experiment 1 demonstrating this in the written modality, and Experiment 2 in speech. Experiment 3 shows that participants are biased towards iconicity when spontaneously generating words, thus clarifying the dynamics of the emergence of iconicity in Experiments 1 and 2 by suggesting that it is a result of production as well as learning biases.

## Experiments

We focus on two types of iconicity that are well documented: sound-shape iconicity (aka kiki-bouba iconicity, extensively explored in Chapters 2 and 3) and duration of motion iconicity. These are particularly interesting because they go beyond mere sound-sound correspondences, and capture basic properties of objects (shape) and actions (duration of motion).

As noted above, many studies present evidence for a role of sound-shape iconicity in word learning, and processing. However, most of these studies confound rounded-round and spiky-spiky associations, by using binary forced choice paradigms (as in the classic kiki-bouba experiments), potentially masking differences in strength between the two associations (see Chapters 2 and 3). Moreover, they studies do not address whether speakers create iconic vocabularies. Perlman, Dale, and Lupyan (2015) present evidence that people creating novel vocalizations heavily utilise iconicity. However, in that study, communication was limited to expressing contrastive dimensions (e.g., small vs. big); vocabulary transmission was restricted to pairs of people intentionally establishing a new communication system, and receiving feedback and correction rather than producing spontaneously. Finally, speakers produced not words but broader vocalizations where iconicity was carried by prosody rather than wordform. It is well known that features like pitch and voice quality can be used iconically by both humans and nonhuman species (Ohala, 1994), but the potential for iconicity in the wordform (i.e. segmental information) is unexplored. Thus these studies represent a new contribution to the literature on this form of iconicity.

Motion iconicity is the extension or reduplication of parts of words to signify repeated action or motion. For instance, in Japanese sound symbolism repetition of syllables can indicate repetition of actions or events, e.g. *goro* = heavy object rolling, *gorogoro* = heavy object rolling repeatedly (Vigliocco & Kita, 2006). In the Niger-Congo language Siwu, the vowel in *kããa* - meaning “looking attentively” - can be greatly extended, and *tsòkwε-tsòkwε* - for “sawing movement” – can be repeated indefinitely, in each case signifying extended duration or repetition (Dingemanse, 2011). While many studies have investigated shape iconicity, very few studies have addressed motion iconicity, despite its frequency in sound-symbolic languages. The major exception is Cuskley (2013), who found that participants recruited for an English language experiment on Amazon Mechanical Turk (and thus presumably not speakers of sound-symbolic languages) were sensitive to motion iconicity, matching words that reduplicate words to stimuli moving at a faster speed. However this sensitivity has not been tested in a production task, making this an important test of speakers of a non-sound symbolic language’s sensitivity in that context.

## Experiments 1 and 2: Iconicity in Language Change

The iterated learning model (Scott-Phillips & Kirby, 2010) approximates cultural evolution using diffusion chains, in which a succession of participants (generations) each learns from their predecessor. An initial participant learns a ‘language’ of novel words for visual stimuli (this language is “generation 0”), and is later tested on these names (unbeknownst to them) for similar novel stimuli. Crucially, testing unseen items compels innovation. Hence when the first generation’s responses, including mistakes and changes, are taught to the second, the language evolves. This process

is repeated between the second and third generation, the third and fourth, etc. Participants remain unaware of being in a chain and believe they are in a simple learning experiment.

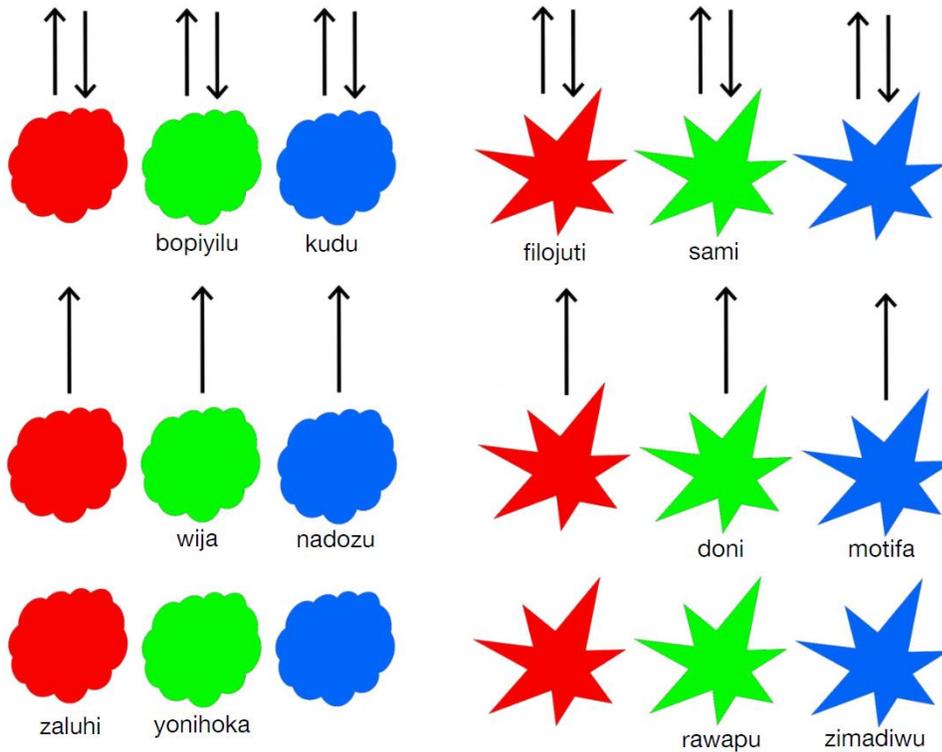
If the changes participants make render the language more learnable or reflect a consistent bias, they should be retained across generations (Christiansen & Chater, 2008). If iconicity meets these conditions, iconic mappings should emerge from an arbitrary initial language and be maintained through generations. We test these predictions in Experiment 1 and 2, across ten generations with a non-iconic initial language. In Experiment 1 participants were taught written words, and in Experiment 2 spoken words. Otherwise materials were the same.

## **Experiment 1**

### **Methods**

**Participants** Sixty native speakers of English (32 women,  $M = 26.3 \pm 8.5$  years old) recruited from the UCL subject pool. This number was chosen to match the number of observations reported in Kirby, Cornish, and Smith (2008). Previously eleven participants were excluded, five for generating fewer than five word types, two for not being monolingual English speakers, one for noticing testing of unseen items, one for misunderstanding when to type in responses, one for reporting memory problems, and one for having been run on the wrong procedure.

## Materials



*Figure 4.1: Stills from the stimuli for Experiments 1 & 2. Arrows denote motion. No arrows: still; one arrow: single motion (c. 0.6s); two arrows: repeated bouncing motion (c. 5s). Stimuli in the original teaching set (“generation 0”) appear above their original names. Stimuli without names were not taught in the first generation, only tested.*

**Visual Stimuli** Eighteen 5s video stimuli, varying on dimensions of shape (round vs. spiky), color (red, green, or blue), and motion (still, single upwards stroke, or up-down bounce – see Figure 4.1). Color was not of interest, but generated enough stimuli to make recalling names challenging (cf. Kirby et al., 2008). Shapes maximized contrast between round and spiky. In Figure 4.1, arrows represent motion (no arrow: still, motion = 0s; single arrow: single upwards stroke, motion = c. 0.6s; dual up-down arrows: repeated bouncing, motion = 5s, duration of individual strokes = c. 0.6s).

**Labels** For the initial language (aka generation 0), iconically neutral letter strings were constructed from the LetterScore norms. Neutral names were generated by combining syllables whose scores summed to c. zero (e.g. one round = mo, one spiky = ti, one neutral = fa: motifa). This ensured that though initial names were neutral, participants had a phonologically varied language. Name length was randomized between two, three, and four syllables.

**Apparatus and Procedure** The experiment used E-Prime 2.0. The procedure closely followed Kirby, Cornish, and Smith (2008, Experiment 2). Participants were told they would learn an ‘alien language’ of word-video pairings. Each language comprised 18 pairings, divided (participant by participant) into a SEEN set (12 items) and an UNSEEN set (six items). The SEEN set was chosen to minimize multiple videos with the same name, to prevent the language ‘collapsing’ into a single word (cf. Kirby et al., 2008). Within this constraint, selection was randomized. Participants were trained on the SEEN set only, but (unbeknownst to them) tested on both sets, to force innovation. A post-experiment questionnaire confirmed that participants typically did not notice the novel items.

Participants learned in three rounds of training, with breaks between rounds. Each round was followed by a testing phase. In each round participants were trained on the SEEN set in two randomized orders. The first frame of each video was displayed for 1 second before the letter string was displayed below the video. The video + name were visible together for 5 seconds.

In each testing phase, participants were presented with videos, and typed names using a standard keyboard. There was no time limit. The first round’s test phase contained only half the SEEN set and half the UNSEEN set, with the second

containing the other half of each set. The final test phase featured all items. Each video was tested a maximum of once per testing phase. The responses in the final test phase became the next generation's training set.

Participants were assigned to one of 6 diffusion chains of 10 generations each. The first participant in each was trained on the initial language. Subsequent generations were trained on a SEEN set drawn from the output of the previous participant in the chain. Participants were unaware of being situated in a chain with other participants

**Design** Our predictors were generation (0 – 10), shape (round vs. spiky), and motion (still vs. single upwards stroke vs. bouncing). For shape iconicity, the dependent variables were LetterScore and WordScore. WordScore was obtained by asking an additional group of 18 participants to rate each label in the experiment for roundness vs. spikiness on a seven-point scale. As with LetterScore, zero was the centre point, with positive = spiky, and negative = rounded. For motion iconicity, the DV was length in letters.

## **Statistical Models**

**Variable Coding** Generation: linear terms indicate the overall directional trend; quadratic terms modulate a linear trend so that we can capture whether the trend slows down in later generations. Such a modulation is plausible, assuming that iconicity emerges and then stabilizes (we had no predictions justifying difficult-to-interpret higher order terms). Shape and Motion were each coded with a single variable. Shape: -0.5 = rounded, +0.5 = spiky. Motion : -1 = still, 0 = single upwards stroke, 1 = repeated motion. Our prediction was that length should correlate with duration of motion (i.e. bouncing stimuli > single stroke stimuli > still stimuli), and this

prediction is built into our single linear code. This coding serves as a conservative way to test our hypothesis, avoiding multiple contrast codes and multiple comparisons.

**Model Construction** We employed mixed-effects models with nested random effects for participant (we deemed the initial language in each chain a 'participant') and diffusion chain, using restricted maximum likelihood estimation. The models were implemented using the package lme4, version 1.7 (Bates, Maechler, Bolker, & Walker, 2014) running in R version 3.1.2 (R Core Team, 2014). In addition to random intercepts for generations and chains, we also included random slopes. We aimed for a design-driven maximal random effects structure, but as the number of observations per statistical unit in the random effects structure was comparatively small, we were limited in the number of slopes we could fit for each analysis. However, we fit slopes for the respective variables our hypotheses claim would be crucial predictors for each particular analysis (see Barr, Levy, Scheepers, & Tily, 2013) - for our analysis of LetterScore and WordScore: shape (by participant) and shape, generation, and the generation $\times$ shape interaction (by chain); and for our analysis of length: motion (by participant), and motion, generation, and the generation $\times$ motion interaction (by chain). We also fit random intercepts for both chains and participants in all analyses.

Our policy in exploring each dependent variable was to begin by constructing an omnibus model that combined predictors in a factorial design up to and including three-way interactions, and then to remove non-reliable interactions and the higher order terms of quadratic generation (if non-reliable) and report the resulting reduced model. Other than quadratic generation, main effects were always retained, whether they were reliable or not.

Confidence intervals were estimated on the basis of the model parameters' standard error values.

## Results

**Learnability** This is the first experiment to be carried out to investigate iconicity using the iterated learning paradigm, so as a “manipulation check” we assessed whether speakers learned, and whether the artificial language became more learnable over the course of the experiment. Following Kirby, Cornish, & Smith (2008) we measured transmission error between generations operationalized as Levenshtein edit distance, a transform-distance index of learnability useful as a manipulation check. We then normalized it by dividing the absolute value by the largest logically possible (i.e. the length of the longer string).

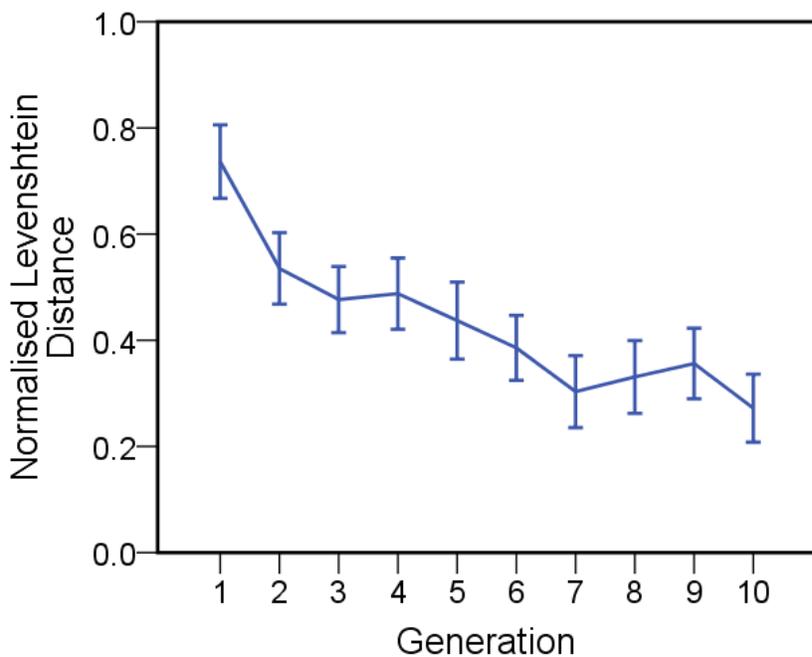


Figure 4.2: Transmission error in Experiment 1. Error bars show 95% confidence intervals.

In the initial omnibus model, the only reliable interactions were between shape and generation, and between generation and manner of motion. All other interactions were non-reliable ( $|t| < 0.7$ ), therefore interactions between other combinations of predictors were removed from the next model (see Appendix 4.2 for all final model specifications for Experiments 1 and 2). In the new model there were reliable main effects of shape ( $\beta = -0.071$ , 95% CI [-0.131, -0.011],  $t = -2.30$ ), indicating that rounded stimuli tended to enjoy lower error rates, and of motion ( $\beta = 0.039$ , 95% CI [0.002, 0.076],  $t = 2.08$ ), indicating that stimuli that move for less time tended to enjoy lower error rates than longer-moving stimuli. Crucially, both linear and quadratic polynomials of generation were also reliable predictors of normalized error. The linear trend was negative ( $\beta = -0.041$ , 95% CI [-0.059, -0.023],  $t = -4.58$ ), indicating an overall decline in error rates across the generations, while the quadratic trend was positive ( $\beta = 0.005$ , 95% CI [ $2.40 \times 10^{-4}$ , 0.010],  $t = 2.06$ ) indicating that improvement in error rates slows in later generations (see Figure 4.2). Moreover there were reliable interactions between shape and quadratic generation ( $\beta = 0.007$ , 95% CI [0.001, 0.012],  $t = 2.45$ ), and motion and quadratic generation ( $\beta = -0.004$ , 95% CI [-0.007, -0.001],  $t = -2.32$ ). These indicate faster or slower reductions in error rate in later generations for different shapes and motions (possibly due to more iconicity emerging for some categories of stimuli, or due to iconicity driven reductions in name length contributing to memorability in certain categories). However, there are no interactions between shape or motion and linear generation, indicating that there was reduction in error across the board.

**Shape Iconicity** was assessed using LetterScore, a syllable-norm based metric; and WordScore, a metric based on direct ratings of “roundness” and “spikyness” of the words produced in the experiment in order to confirm robustness of our findings, and for direct comparability to Experiment 2.

**LetterScore** ratings were obtained using the LetterScore norming mentioned in Chapter 3. All consonant-vowel pairings possible in English orthography featuring consonants with only one canonical pronunciation (N=85; c, g, q, and x were excluded) were rated by monolingual English speakers who did not participate in the other studies (N = 28, 12 women,  $28.5 \pm 12.0$  years old) on a ten-point scale anchored by a circle (1) and a star (10) (see Appendix 4.1 for the scale). A centred scale was created by redefining the mean rating (5.04) as zero. Each letter’s LetterScore was defined as the mean of the ratings of the syllables it appeared in (see Appendix 3.2 for a complete list of syllable ratings). A letter that tended to appear in spiky-sounding syllables would receive a positive (spiky) score (e.g. z = 1.06), a letter that appeared in round-sounding syllables would receive a negative (round) score (e.g. m = -1.46). By defining a word’s LetterScore as the mean of its letters’ LetterScores, we obtained an index of the iconicity of written words (four consonants not featured in the original norming receive a score of zero). Positive scores represent spikiness, and negative scores roundness. See Table 4.1 for a list containing each letter’s LetterScore.

Letter	LetterScore	Letter	LetterScore
k	1.60	d	-0.25
z	1.20	s	-0.37
i	1.00	w	-0.45
t	0.94	u	-0.54
v	0.48	h	-0.56
e	0.47	b	-0.70
j	0.13	l	-0.79
p	0.12	o	-1.01
f	0.10	m	-1.33
a	0.08	g	0.00
y	-0.01	q	0.00
r	-0.04	x	0.00
n	-0.07	c	0.00

*Table 4.1: LetterScores, arranged in descending order of spikiness. Consonants exclude from the analysis appear highlighted in red.*

Comparing the LetterScore consonants to the mean Syllable Scores for the phonemes those letters represent (see Chapter 2, English vowels are too phonetically polyvalent for such a comparison to be possible) I found that a simple regression gave  $r^2 = .69$ , indicating that LetterScore norms correlate well with the speech based norms we reported earlier.

This metric was used to assess the roundness and spikiness of words in Experiment 3 as well as Experiment 1.

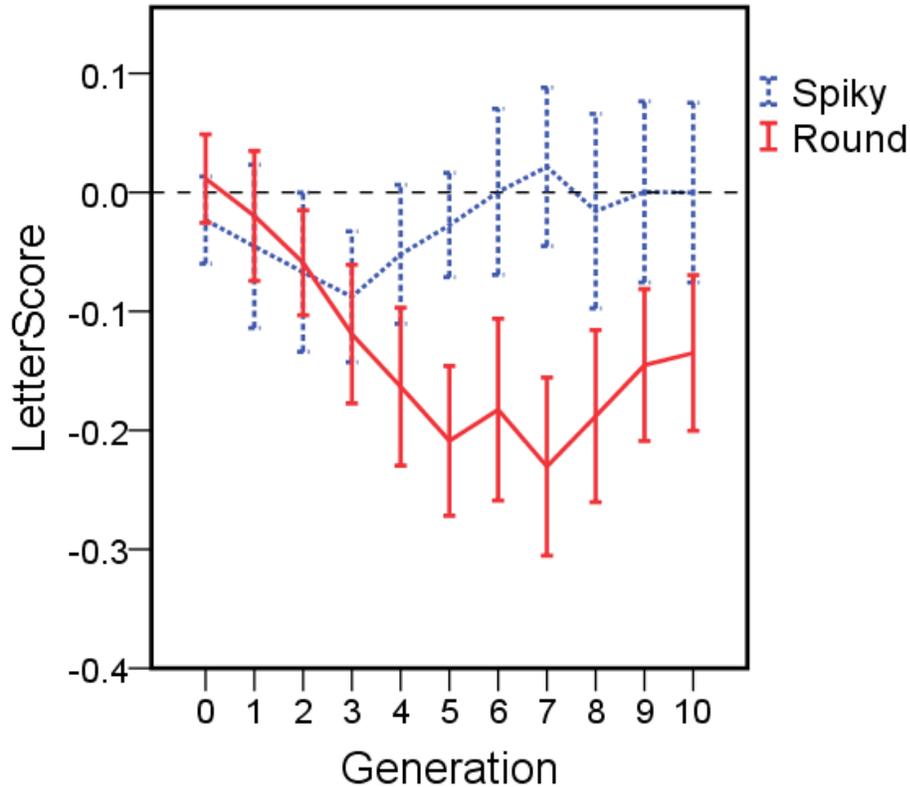


Figure 4.3: The interaction of shape and generation in Experiment 1, measured using LetterScore. Error bars represent 95% confidence intervals.

For LetterScore the initial omnibus model failed to converge, therefore we removed quadratic generation from the random effects term (as between-chain variance for this was very low). This new omnibus model did converge, and showed that the only reliable interactions were between shape and generation; all others were unreliable ( $|t| < 1.4$ ). Unreliable interactions were removed from the final model. The final model contained reliable effects of quadratic generation ( $\beta = 0.003$ , 95% CI [0.001, 0.005],  $t = 3.22$ ), indicating an overall u-shaped trend, and shape ( $\beta = 0.155$ , 95% CI [0.066, 0.243],  $t = 3.41$ ), indicating that across generations round shapes tended to have lower LetterScores than spiky shapes. Crucially, shape interacted with linear generation ( $\beta = 0.025$ , 95% CI [0.007, 0.043],  $t = 2.78$ ), indicating that names for

round and spiky shapes diverged in the expected direction over the generations (see Figure 4.3). The interaction between shape and quadratic generation was also significant ( $\beta = -0.005$ , 95% CI [-0.009, -0.002],  $t = -3.34$ ), indicating that this divergence slowed in later generations.

Inspection of Figure 4.3 suggests that these interactions are driven by round stimuli's names becoming rounder, with spiky stimuli remaining closer to neutrality. To assess this we ran separate post-hoc analyses for round and spiky stimuli including only linear and quadratic generation as fixed effects. For round stimuli, both linear generation ( $\beta = -0.018$ , 95% CI [-0.034, -0.001],  $t = -2.12$ ) and quadratic generation ( $\beta = 0.006$ , 95% CI [0.003, 0.008],  $t = 4.72$ ) were reliable. However neither were reliable for spiky stimuli ( $|t| < 1$ ).

WordScore Norming: Three participants were excluded prior to analysis for routinely answering questions in less than 0.5s, indicating lack of engagement with the task. This left 18 native English speaking participants (age  $M = 36.9 \pm 11.2$ ; nine women) to rate each label in the experiment for roundness vs. spikiness using a seven point scale anchored by round and spiky shapes from Experiment 3 (see Appendix 4.1 for the scale). This norming study was carried out as an online survey using the online testing platform Testable (2016). For half of participants 1 was assigned a round shape and 7 a spiky shape, for the other half this was reversed. After ratings were collected those collected using the round-high scale were 'flipped' such that 7 became 1, and 1 became 7 etc.. This meant that for all data 1 represented the roundest rating, and 7 the spikiest. The mean of each token's ratings was then taken. This was its WordScore.

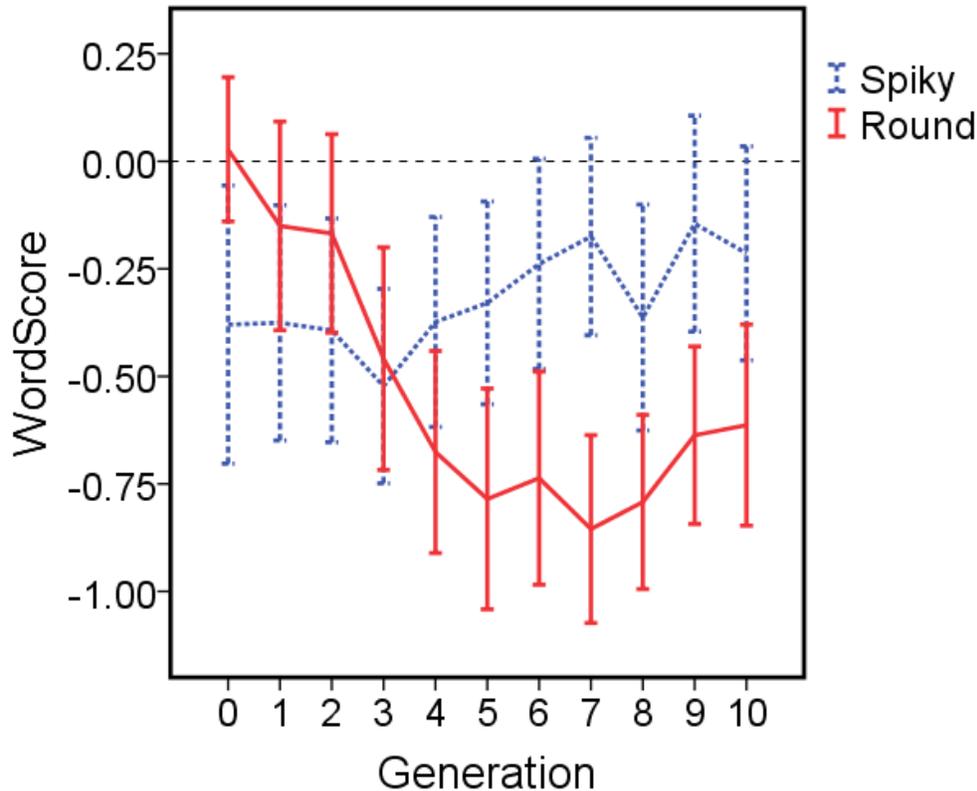


Figure 4.4: The interaction of shape and generation in Experiment 1, measured using WordScore. Error bars represent 95% confidence intervals.

Using WordScore, the initial omnibus model did not feature significant interactions between shape and motion, or between shape, motion, and generation ( $|t| < 1.5$ ). Unreliable interactions were removed from the final model. In the final model there were reliable effects of quadratic generation ( $\beta = 0.010$ , 95% CI [0.004, 0.017],  $t = 3.01$ ), and shape ( $\beta = 0.3813$ , 95% CI [0.1647, 0.598],  $t = 3.45$ ). Somewhat unexpectedly, we found an interaction between linear generation and motion ( $\beta = 0.039$ , 95% CI [0.020, 0.057],  $t = 4.05$ ), indicating that moving stimuli tended to acquire spikier names over the generations, possibly inked to the spiky trajectories of their motion, whereas still stimuli acquired rounder names, possibly due to the fact that round words also tend to suggest slowness – Perniss, Thompson, & Vigliocco

(2010). Crucially we again found a reliable interaction between shape and linear generation ( $\beta = 0.098$ , 95% CI [0.058, 0.138],  $t = 4.76$ ), modulated by a negative interaction between shape and quadratic generation ( $\beta = -0.0173$ , 95% CI [-0.0283, -0.006],  $t = -3.077$ ). See Figure 4.4.

As with LetterScore, we analysed round and spiky stimuli separately. Again, for round stimuli both linear generation ( $\beta = -0.073$ , 95% CI [-0.141, -0.005],  $t = -2.12$ ) and quadratic generation ( $\beta = 0.019$ , 95% CI [0.011, 0.027],  $t = 4.65$ ) were significant predictors. Again, neither predictor was significant for spiky stimuli ( $|t| < 1$ ).

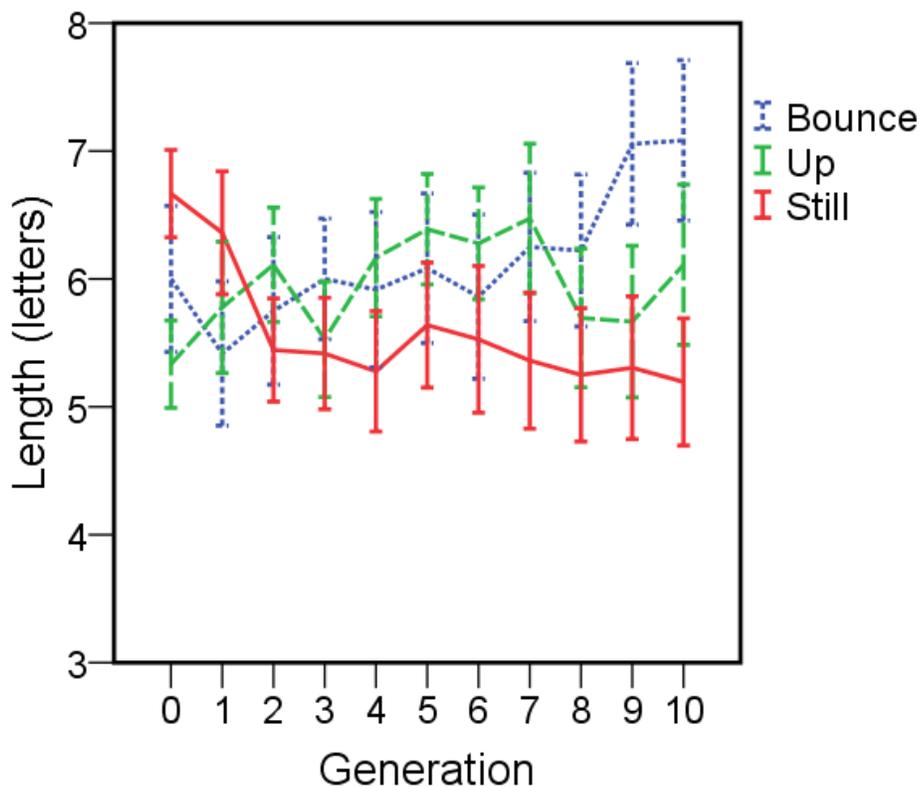


Figure 4.5: The interaction of motion and generation in Experiment 1, measured using length in letters. Error bars represent 95% confidence intervals.

Motion Iconicity was measured by number of letters per name; again we included Generation 0 to reflect the initial vocabulary. The initial omnibus model's only reliable interaction was between motion and linear generation; all others were unreliable ( $|t| < 1.5$ ) and were therefore excluded from the final model, along with quadratic generation, also unreliable ( $|t| < 1$ ). The final model showed a reliable main effect of shape ( $\beta = -0.344$ , 95% CI [-0.521, -0.167],  $t = -3.80$ ), indicating that rounded shapes tended to have longer names (conceivably because systematic mappings evolved linking rounded shapes to longer names, cf. the emergence of morphology in Kirby, Cornish, & Smith, 2008). Crucially, it also showed a reliable interaction between motion and linear generation ( $\beta = 0.123$ , 95% CI [0.049, 0.197],  $t = 3.25$ ), indicating that duration of motion predicted change in length of name over the generations (Figure 4.5).

To further investigate the interaction between motion and linear generation, we fit separate models for each motion, with linear generation only as fixed effect. For still stimuli linear generation was a reliable predictor ( $\beta = -0.116$ , 95% CI [-0.204, -0.028],  $t = -2.59$ ) indicating that names for still stimuli became shorter across generations. There was no reliable effect of linear generation ( $|t| < 1$ ) for stimuli in the single upwards stroke condition. For bouncing stimuli there was a reliable effect of linear generation ( $\beta = 0.134$ , 95% CI [0.030, 0.237],  $t = 2.53$ ), indicating that these names became longer over generations.

Similar results obtain for syllable-based analysis (see Appendix 4.3).

Thus to summarize, iconicity emerges for both shape and motion.

## Experiment 2

Experiment 2 is a replication of Experiment 1 but with spoken rather than written names.

### **Methods**

**Participants** Sixty native speakers of English (41 women,  $M = 22.6 \pm 8.2$  years old) recruited from the UCL subject pool, all of whom received cash or course credit as payment. Six participants were excluded, two for generating fewer than five word types, one for not being a monolingual English speaker, and three because of problems in recording their responses.

**Materials, Apparatus, Procedures and Design** Were as Experiment 1, with the following exceptions: (1) the initial language (generated as in Experiment 1) was recorded according to its most obvious pronunciation by a North American linguist. (2) labels were presented via headphones. (3) Participants produced labels into a microphone, and heard names through headphones. Headphones: Beyerdynamic DT100; microphone: Audio-technica ATR20. Between generations names were edited to remove silence, and volume was normalized.

See Table 4.2 for the initial language.

Shape	Colour	Motion	Name
Rounded	Blue	Still	[hɑɪ'maɪwaɪ]
Rounded	Blue	Up	[ni'lʊvaɪ]
Rounded	Green	Bounce	[kiwəʊ'mʊdʒaɪ]
Rounded	Green	Up	[dəʊpaɪ'həʊkɑ]
Rounded	Red	Bounce	['wʊzɑ]
Rounded	Red	Still	['təʊfɑ]
Spiky	Blue	Bounce	[zɪdʒəʊ'taɪjəʊ]
Spiky	Blue	Up	['zaɪməʊ]
Spiky	Green	Bounce	[səʊti'kaɪbəʊ]
Spiky	Green	Still	[faɪ'mɑnɑ]
Spiky	Red	Still	[ləʊjaɪ'pu]
Spiky	Red	Up	[hɑ'sɑ]

*Table 4.2: The initial language for Experiment 2*

## Results

### Shape Iconicity

Norming: 101 native English speakers ( $M = 32.4 \pm 9.7$ , 41 women) recruited through the website Prolific Academic were each given slightly more than 100 speech tokens to rate, meaning that each of the 1092 speech tokens from the experiment was rated ten times on average. The study was performed online on Qualtrics (2015). Rating worked in the same way as for Experiment 1.

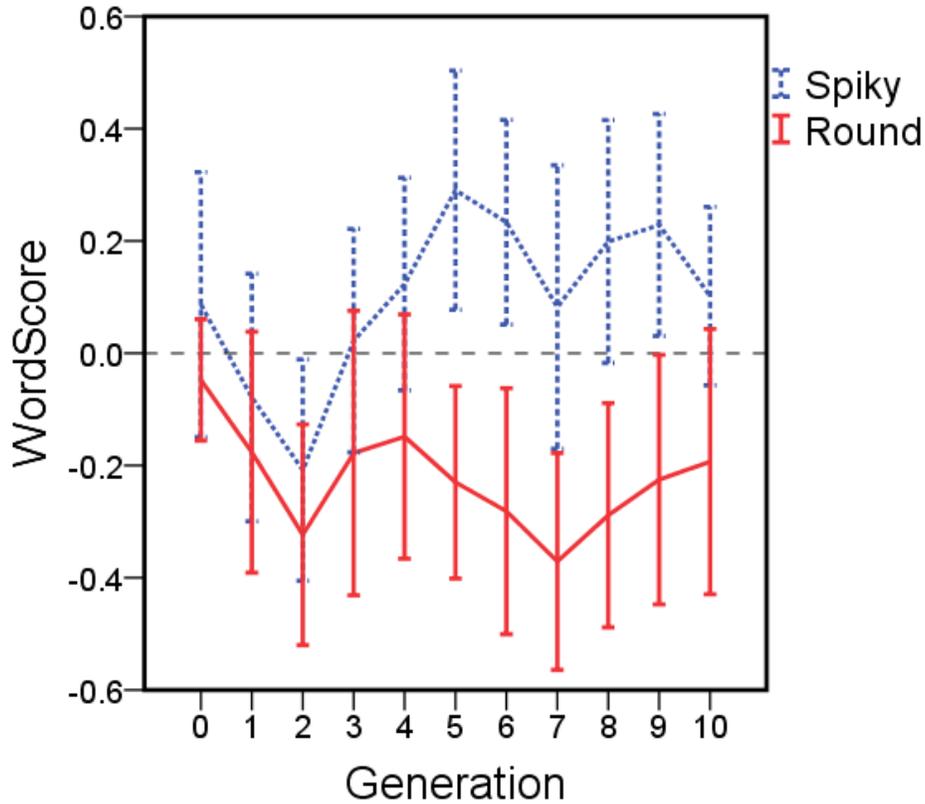


Figure 4.6: The interaction between shape and generation Experiment 2, measured using WordScore. Error bars represent 95% confidence intervals.

The initial model featured no terms suitable for removal. There was a reliable interaction between motion and quadratic generation ( $\beta = 0.008$ , 95% CI [0.002, 0.015],  $t = 2.55$ ), indicating that motion modulated the rate of change across the generations; a reliable interaction between shape and motion ( $\beta = -0.280$ , 95% CI [-0.425, -0.134],  $t = -3.76$ ), indicating that the difference between round and spiky stimuli was stronger for stimuli that moved less; and a difficult to interpret interaction between quadratic generation, motion, and shape ( $\beta = 0.014$ , 95% CI [0.001, 0.027],  $t = 2.14$ ). Importantly, there was a reliable effect of shape ( $\beta = 0.438$ , 95% CI [0.018, 0.856],  $t = 2.04$ ), indicating that round stimuli tended to have rounder names than spiky stimuli. Inspection Figure 4.6 suggests this is a main effect rather than an

interaction with generation because iconicity emerges quite suddenly, and is not well captured by a linear trend.

As in Experiment 1, we analyzed round and spiky stimuli separately with fixed predictors as linear and quadratic polynomials for generation. The round model featured a reliable intercept only ( $\beta = -0.261$ , 95% CI [-0.489, -0.033],  $t = -2.24$ ), indicating that round stimuli tended to have round names. The spiky model featured no reliable predictors ( $|t| < 1.6$ ). Thus as in Experiment 1, iconicity seems to target roundness. We return to this point in the discussion.

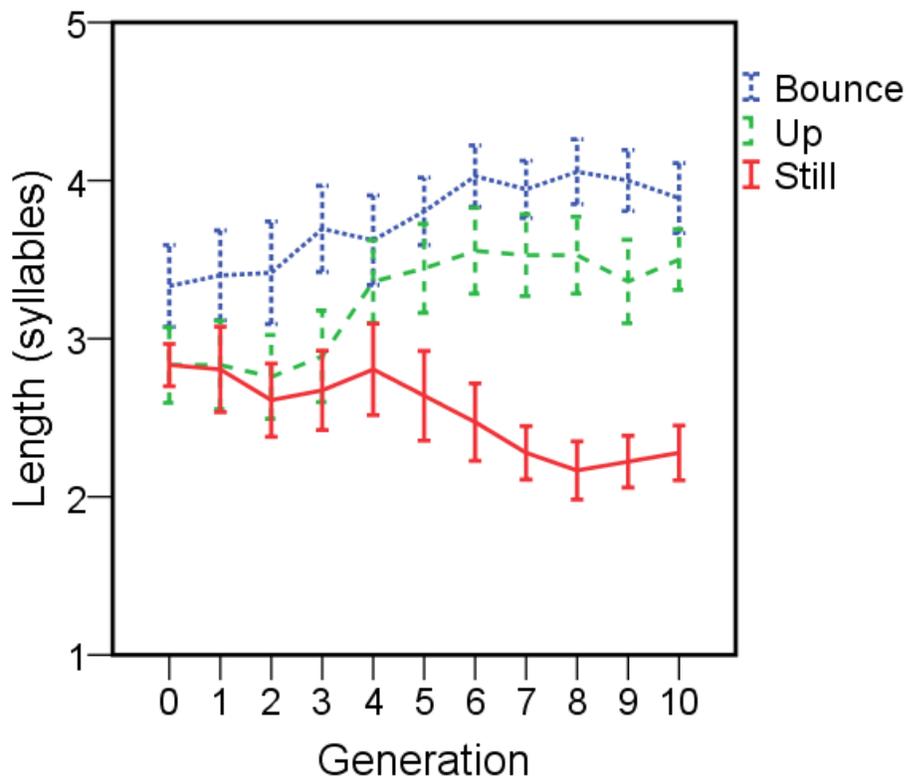


Figure 4.7: The interaction of motion and generation in Experiment 2, measured using length in syllables. Error bars represent 95% confidence intervals.

**Motion Iconicity** was measured as length in syllables (coded by the researchers).

The initial omnibus model's only reliable interaction was between motion and linear

generation, all others were unreliable ( $|t| < 1.5$ ), and were excluded from the final model. In the final model we found a reliable main effect of quadratic generation ( $\beta = -0.010$ , 95% CI [-0.017, -0.004],  $t = -3.08$ ), indicating that, overall, length increased over earlier generations and decreased over later ones; a main effect of shape ( $\beta = 0.092$ , 95% CI [0.012, 0.172],  $t = 2.24$ ), indicating that spiky shapes tended to have longer names than round ones (perhaps the result of systematic mappings, as in Experiment 1); and a main effect of motion ( $\beta = 0.665$ , 95% CI [0.435, 0.894],  $t = 5.68$ ), indicating that stimuli that moved more had longer names. Crucially there was a reliable interaction between motion and linear generation ( $\beta = 0.073$ , 95% CI [0.027, 0.118],  $t = 3.14$ ), indicating that duration of motion predicted change in name length over the generations (see Figure 4.7).

As in Experiment 1, we investigated the interaction between motion and linear generation by fitting separate models for each motion (both linear and quadratic generation). For still stimuli only linear generation was a reliable predictor ( $\beta = -0.069$ , 95% CI [-0.115, -0.023],  $t = -2.92$ ), indicating that the names for still stimuli tended to get shorter over the generations. For stimuli that described a single upwards stroke, the only reliable effect was quadratic generation ( $\beta = -0.018$ , 95% CI [-0.031, -0.005],  $t = -2.76$ ), suggesting that length tended to grow in earlier generations, and that this trend was reversed in later generations. Finally, the model for bouncing stimuli featured effects of both linear ( $\beta = 0.076$ , 95% CI [0.005, 0.147],  $t = 2.092$ ) and quadratic generation ( $\beta = -0.012$ , 95% CI [-0.021, -0.003],  $t = -2.53$ ), indicating that length grew overall through the generations, but that this trend slowed in later generations. In summary, change in length of name once again correlates with duration of motion.

Thus once again, iconicity emerged for both shape and motion.

Both Experiment 1 and 2 show that iconicity emerges for shape and motion in a model of language change. However, it is not clear what dynamics underlie this. As discussed in the introduction, there are at least two mechanisms by which cultural evolution could create fitter languages: natural selection in the classical sense (which crucially involves competition between units of selection: words in this case), and what I term *selective mutation*, whereby copying errors between one agent and the next tend to drive the language towards learnability without competition.

It is not clear what the balance is between these mechanisms in this case. We can envision several different scenarios: perhaps when naming unseen items (i.e. items that have not been trained) in the testing phase, participants preferentially extend iconic rather than non-iconic words from their training set. In this case word would directly compete with each other on the basis of iconicity. Alternatively participants might just make fewer errors when recalling iconic vocabulary, meaning that wordforms that are noniconic tend to mutate rapidly until by chance they become more iconic and thus more stable. This would be an instance of selective mutation. Yet another possibility is that selective mutation operates not through learnability biases but through error biases: if participants' are just as likely to make recall errors for both iconic and noniconic vocabulary, but produce errors that tend to introduce iconicity, then each generation the language will be topped up with iconic forms, ensuring that for the first few generations at least the amount of iconicity present increases.

To begin to address this, I ran one further study, Experiment 3, where participants were asked to spontaneously generate words for visual stimuli differing on shape

and motion. As discussed in the introduction, iconicity in production is little studied so far. However, if Experiment 3's spontaneously generated words feature iconicity, this increases the plausibility of participants in Experiment 1 and 2 being biased towards introducing iconicity in the testing phase, meaning that at least part of the iconicity that emerges is driven by selective mutation.

### **Experiment 3: Iconicity in Names for Novel Objects**

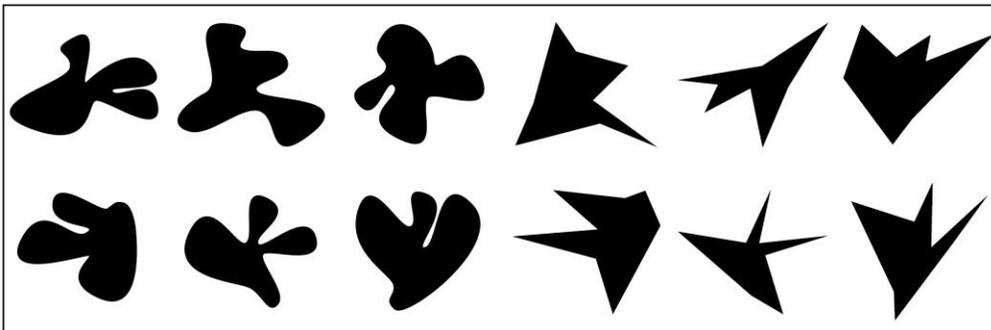
In Experiment 3, we ask whether modern speakers accustomed to using a sophisticated and arbitrary system of communication (i.e. English) nonetheless spontaneously resort to iconicity when learning and generating new labels. Because of the difficulty of creating new labels outside of the phonotactic constraints of a native language (though see Perlman, Dale, & Lupyan, 2015), we address this question by eliciting (written) productions of possible words.

If language users are biased towards iconicity, then speakers should create iconic labels for novel objects. Here we showed speakers two novel objects - a rounded or spiky figure (Figure 4.8); and a dot moving once or repeatedly. We asked them to type a new word for the object. By presenting a single object for each dimension of interest (shape or motion) rather than contrasting objects, we avoid highlighting our dimension of interest (as previous studies did), and therefore do not lead responses. For shape, we measured iconicity using norms for goodness of fit to a round or spiky figures of each letter, obtained from separate participants. For motion, we counted letters.

## Methods

**Participants:** 97 first-year undergraduates at University College London participated as part of a laboratory class (age  $M = 19.0 \pm 0.8$ ), 83 of them women, 63 native Anglophones. Of these 97, we excluded 45 who – in spite of explicit and repeated instructions to avoid creating names based on English words – created names based on English words (e.g., ‘blob’ for a rounded shape, and ‘angulary’ for a spiky shape). Thus, we only included data from the remaining 52 participants in the analyses (see Appendix 4.4 for Anglophone only analyses).

## Materials



*Figure 4.8: Examples of round and spiky shapes used in Experiment 3 (stimuli for the motion-iconicity condition were a dot moving up and down one vs. four times).*

**Visual Stimuli** Each participant saw one stimulus from a set of 16 spiky or 16 rounded shapes, followed by one of two motion-targeted stimuli, each a video of a small black moving dot. One showed a dot making a single upwards stroke, the other a dot in an up-down-up-down motion. Each video lasted about six seconds (single stroke: 1.25s, repeated stroke: 5s).

The spiky shapes were generated by a Matlab 2012a randomization script based on the procedure reported in Monaghan et al. (2012). The round stimuli were generated

by using the GNU image manipulation program (GIMP, 2012) to smooth the spiky shapes' sides into Bezier curves with their fixed points on the spiky shapes' angles, and then matching for size. Stimuli were 600\*600 pixels images comprising the shape in black on a white background.

**Apparatus and Procedure** The participants had to create a name for each stimulus by typing a response. The experiment was run using E-Prime 2.0.

**Analysis of sound-shape iconicity** LetterScore ratings were obtained using the LetterScore norming outlined in the methods for Experiment 1.

## **Results**

Data from 52 participants were analysed. Independent variables were round vs. spiky shape, and single vs. repeated motion; names produced for shape and motion stimuli were analysed separately using independent-samples tests. Dependent variables were: LetterScore (ratings of "roundness" vs. "spikiness" of syllables, see Methods) ranging from -5 to 5), and length of name in letters and syllables.

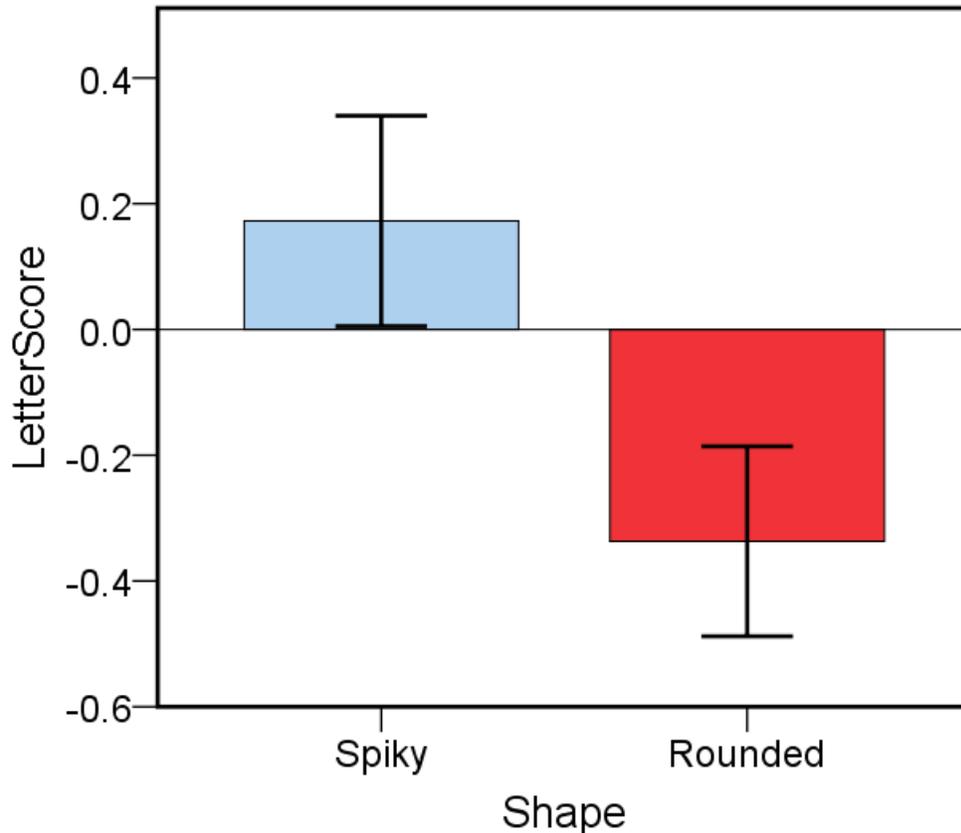


Figure 4.9: The relationship between shape and shape iconicity in Experiment 3. Error bars represent 95% confidence intervals.

**Shape Iconicity** LetterScore significantly differed between the rounded ( $n = 26$ ,  $M = -0.337$ , 95% CI [-0.488, -0.186],  $SD = 0.374$ ) and spiky conditions ( $n = 26$ ,  $M = 0.173$ , 95% CI [0.005, 0.340],  $SD = 0.414$ ): with  $t(50) = 4.65$ ,  $p < .001$ ; difference = 0.509, 95% CI [0.290, 0.729], Cohen's  $d = 1.29$ . See Figure 4.9. Both the rounded ( $t(25) = -4.59$ ,  $p < .001$ ,  $d = 0.901$ ) and the spiky ( $t(25) = 2.12$ ,  $p = .044$ ,  $d = 0.418$ ) conditions are significantly different from zero (Figure 4.9).

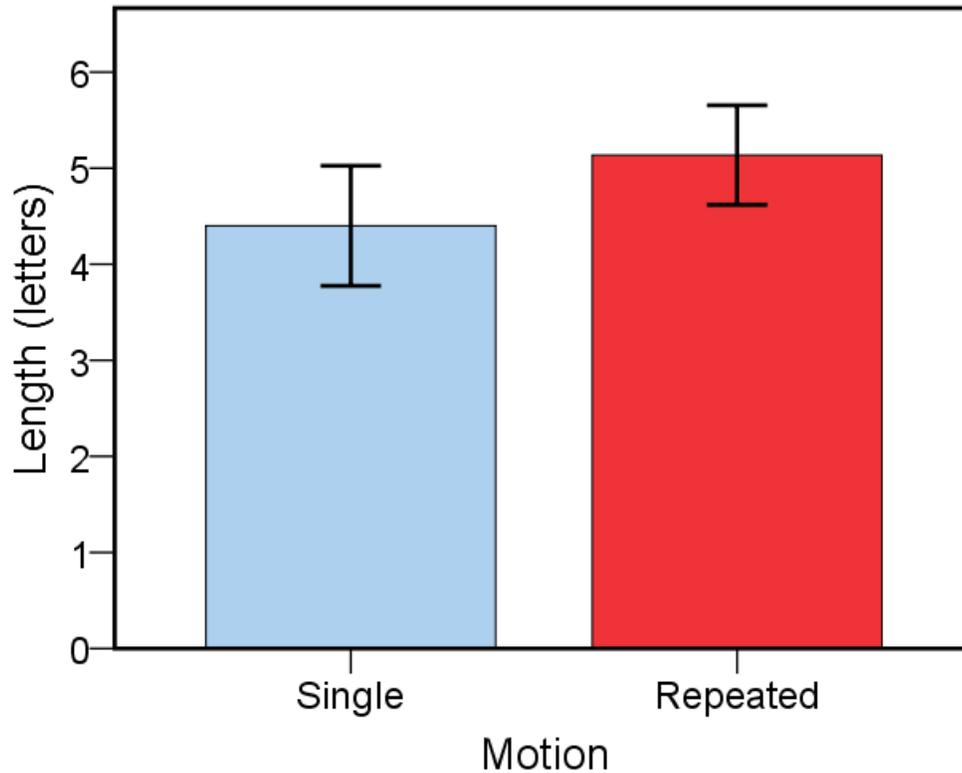


Figure 4.10: The relationship between motion and duration of motion iconicity in Experiment 3. Error bars represent 95% confidence intervals.

**Motion Iconicity** Length significantly differed between the single-motion (median = 4) and repeated-motion conditions (median = 5):  $U = 214.5$ ,  $p = .029$ . See Figure 4.10. Syllable analysis gives similar results: see Appendix 4.5.

## Discussion

We asked whether speakers are biased towards creating and learning words that incorporate iconicity for visual properties of referents (shape and motion), and whether such iconicity would emerge and be sustained in a model of language evolution. In Experiments 1 and 2 we used iterated learning to show that the same

kinds of iconicity spontaneously emerge in a model of cultural evolution. Moreover, for shape iconicity we found an asymmetry: rounded iconicity was stronger than spiky iconicity. Such asymmetry has not been reported before, and may have important implications for the mechanisms of the iconicity, discussed below.

Experiment 3 demonstrated that shape and motion features are salient to speakers, who - when creating new labels for novel objects - spontaneously express them iconically. This suggests that biases in production may be at least partly responsible for the results seen in Experiments 1 and 2.

Experiments 1 and 3 are limited in that there is no language that uses written words as its primary modality, and although writing activates phonology (Rastle and Brysbaert, 2006), iconic correspondences could be represented unimodally in these experiments, mapping text to visual properties of shape or motion. Experiment 2 addresses these shortcomings by replicating Experiment 1 with spoken labels. The results were very similar, except that shape iconicity emerged quickly and thus did not interact with generation.

Importantly, Experiments 1 and 2 suggests that once present, iconicity tends to stay in the lexicon. However, this however does not mean that we should expect iconicity to progressively accrue in lexica forever. As described in the introduction, the pressure towards iconicity is one of several acting on communicative systems, and will eventually be balanced by other pressures (e.g. towards discriminability, Perniss & Vigliocco, 2014). But we would expect some amount of stable iconicity to be present in all languages. The fact that the same iconic dimensions studied here feature among many spoken languages underscores the generalizability of these observations.

## **Iconicity and Combinatoriality**

One of the few precedents for this work is Verhoef, Kirby, and de Boer (2016). Verhoef et al. performed iterated learning experiments where participants had to learn and produce names for pictures of novel mechanical items. However, these names were not in speech or text, but were instead produced using slide whistles. This was because the authors of the study were interested in the emergence of phoneme-like combinatorial structure, and wanted to remove the possibility of participants simply imposing the structure of their native language. Verhoef et al. found evidence that a certain amount of iconicity emerged in their first, basic condition (which worked similarly to Experiment 2 here): for instance, in one case an object with an ascending row of holes was represented using a series of staccato notes rising in pitch. They also included a second condition that made stable iconicity impossible by shuffling the assignment of names to items after each generation. Interestingly Verhoef et al. found that in this second, non-iconic condition, the level of combinatorial structure increased and the level of transmission error decreased faster than in the first condition (though by the eighth and final generation the first condition had 'caught up').

The combinatoriality result suggests that certain kinds of iconicity are incompatible with certain aspects of combinatorial structure (and vice versa): presumably the iconicity initially present in the first condition was expressed holistically, impeding the emergence of combinatoriality. However these results do not in fact suggest that combinatoriality and iconicity are in conflict per se: there was no measurable decrease in iconicity over the generations of the first condition, even though combinatoriality emerged over this time. This suggests that rather than being driven out by combinatoriality, iconicity instead evolves in order to remain consistent with it.

Thus iconicity and combinatoriality can coexist. The findings of the current chapter reinforce this conclusion by showing that iconicity emerges in iterated learning experiments where combinatorial structure is present from the outset (in the form of letters or phonemes).

As noted above, across most generations the second, non-iconic condition appeared to enjoy a learnability advantage. However it would be premature to conclude that this contradicts the argument that iconicity makes vocabulary more learnable. Firstly, the error metric was automated, and as Verhoef et al. note, it may underestimate the subjective similarity between successive iterations of an iconic form (i.e. if the crucial iconic element of a name was a pitch glide that changes direction half way through to represent a V-shaped contour in the image, then participants may perceive successive iterations of the name as very similar, even if they've undergone e.g. mirror transforms that might trick the artificial metric into measuring them as very different). Thus with respect to the crucial elements of the names, error may have been overestimated. Moreover, even if this is not true, then iconicity might still directly benefit learnability while indirectly impeding it: iconicity might help make holistic names more learnable while indirectly limiting how much more learnable names can become by stymying the appearance of combinatoriality. This is perfectly consistent with the claim that all else being equal, iconicity helps learning, and that once a combinatorial system is in place, iconicity within this system helps.

The studies in this chapter crucially extend Verhoef et al.'s results by showing the emergence of iconicity in the physical and combinatorial media used by natural languages. More generally, the interplay between iconicity and combinatorial structure will be an interesting line of enquiry for future research. It may be that apparent differences in amount of iconicity between spoken languages are related to

those languages' phonotactics (i.e. restrictions on the combinatorial structure of words) as regards phoneme inventory and restrictions on and mechanisms for phenomena like reduplication.

### **Iterated Learning vs. other Mechanisms for Introducing Iconicity to the Lexicon**

Our results here provide a proof-of-concept that domain general biases will suffice to explain the presence of iconicity across the world's languages. They do not directly provide evidence against claims like Kita's (2008), that e.g. sound-symbolism is a vestige of an iconic protolanguage, or another conceivable claim: that iconicity is the result of innate linguistic biases specific to infants and (perhaps also) parents. However, on the basis of what we know right now I would argue that such hypotheses are superfluous. The more parsimonious explanation is that domain general biases operating throughout the lifespan combined with language change (both of which we know exist) are responsible for introducing and maintaining iconicity in lexica.

### **Dynamics of Cultural Evolution**

Experiments 1 and 2 show that an initially arbitrary lexicon evolves towards iconicity in iterated learning. However, Experiments 1 and 2 themselves do not resolve exactly how the cultural evolution plays out, and particularly whether it is driven by competing forms, or by what I term selective mutation (as manifest in e.g. production biases).

Experiment 3, however, shows that when participants spontaneously generate names for similar stimuli to those used in Experiments 1 and 2, they incorporate iconicity. This suggests that at least some of the evolution in Experiments 1 and 2 is driven by production biases instantiating selective mutation. However, this is not to say that learning biases and competition between forms are not also in play. In fact given literature on the benefit of iconicity to learning reviewed in Chapters 1 and 2, it is likely that learning biases do play a role. Questions about the dynamics of such cultural evolution; questions about the benefits of iconicity to memory, processing, and referential disambiguation; general understanding of the cultural evolution; and potentially even historical linguistics are likely to be mutually informative in future.

### **Iconicity as a Universal?**

Iconicity only covers a rather small part of the vocabulary even in iconicity-rich languages such as Japanese (Tomasello, 2008; of course, sign languages are considerably richer in iconicity). However, iconicity is not limited to speech. Language is learned and has evolved in face-to-face contexts where speech is accompanied by gestures (and also head and body movements, prosodic modulation, and eye gaze; Kendon, 2004). Communicative utterances comprise speech along with additional visible actions (gesture etc.). From this perspective, iconicity in spoken languages is not limited to speech, but is also expressed in the co-speech gestures (McNeill, 1992), and in prosodic modulation (Shintel et al., 2006; Perlman et al., 2015). Framed this way, spoken language is much richer in iconicity (Perniss & Vigliocco, 2014).

This can be explained on the assumption that iconicity is the result of universal biases amplified through cultural evolution. If this is indeed the case, and if the iterated learning thesis is correct about other features of language, then it potentially places iconicity on a level footing with other universals.

### **Language Origins – Gesture First?**

On a quite different note, these results may have some bearing on language evolution in a distinct sense: the biological origins of language. In gesture-first theories of language evolution, iconicity is argued to have had a central role in the origins of language by bringing objects and events absent from the immediate physical environment to “the mind’s eye”, and thereby establishing reference (Arbib, Liebal, & Pika, 2008; Corballis, 2009; Sterelny, 2012; Tomasello, 2008). Spoken languages are argued to lack potential for iconicity beyond trivial cases (i.e. onomatopoeia, Tomasello, 2008). These results show that this is not the case for iconicity for shape and motion, both of which go beyond mapping sound to sound. Thus these findings question the necessity of assuming that symbolic gestural systems came first in language evolution even if iconicity was crucial to proto-language (though of course does not prove that they did not come first). But is this kind of iconicity a useful means of displacement outside the lab? Some evidence that it is comes from work by Berlin (2006), who found iconicity in the names for animals of different shape and dimensions in a survey of small-scale societies.

## **Mechanisms of Iconicity**

An open question is the cognitive and neural mechanisms underlying iconic mappings such as shape and motion iconicity. As discussed in Chapter 2, it has been argued that shape iconicity is based on cross-wiring between adjacent brain areas, or on cross-modal sound-shape analogies (Ramachandran & Hubbard, 2001). Our results suggest that shape iconicity may be stronger for rounded than for spiky objects. This has not been reported before, because previous studies relied on mutually determining pairs of name-shape mappings. My speculative hypothesis is that the effect is partly driven by similarity between rounded shape and rounded lips in bouba-type words, with no comparable association between lip shape and spiky objects. Thus some types of iconicity may be understood as originating through unimodal phenomena (either within audition, e.g. onomatopoeia, or vision, as with shape iconicity). This however is unlikely to be the whole story (see Chapters 2 and 3). Interestingly, the round-spiky asymmetry seems to be stronger for production tasks, perhaps reflecting greater activation of articulatory representations in these.

Whether motion iconicity has a similar explanation, and what feature of motion it targets requires further work. Duration of motion seems a parsimonious possibility, as this maps straightforwardly onto word length, but other possibilities include repetition of motion, complexity of motion, and presence vs. absence of motion.

# Chapter 5: Comprehension of Iconic Gesture in Language Acquisition<sup>29</sup>

## Introduction

It has been argued that iconicity may play a significant role in language acquisition (Imai & Kita, 2014; Perniss & Vigliocco, 2014). However, as applied to spoken languages like English this claim seems problematic: where's the iconicity?

In fact, when considered in communicative context, young children's experience in any native language involves abundant opportunities for iconic input from a different source, namely co-speech gestures<sup>30</sup>. Co-speech gestures come in several major varieties: deictics (i.e. ostensive gestures like pointing), beat gestures (used rather like prosody to emphasise certain words), conventional gestures (like the thumbs up), and iconic gestures. Like sign language, iconic gesture has the potential for much richer iconicity than speech, being capable of mimicking a vast range of actions and physical forms to communicate both literal and figurative meanings. Moreover, in adult-adult interactions iconic gestures are highly frequent, accounting for about 30% of gesture produced (McNeill, 1992).

How much of a role, then, do iconic gestures actually play in language acquisition, and - in particular - in acquiring the meaning of words? In this introduction I will summarise the literature on several key questions: Do children produce iconic

---

<sup>29</sup> Thanks to Linda B. Smith for contributions to the study concept and design, and to Char Wozniak for recording stimuli.

<sup>30</sup> Some consider speech and gesture to be an integrated system, whereas others maintain that gesture is strictly extra-linguistic. I don't aim to argue for or against either view here, but simply to explore whether children use iconic properties of gestures to learn about word and utterance meaning, regardless of whether gesture is in some sense part of language or just an ancillary cue.

gestures? Do children comprehend iconic gestures (and can they use them to learn about the meaning of words)? What limits children's ability with iconic gestures at early ages? Are there iconic gestures in child-directed communication?

In the experiment that follows I will narrow my remit, and present data on 18-month-olds that bears on questions two and three.

### **Do children produce iconic gestures?<sup>31</sup>**

Much of the work on young children and iconic gesture has focused on production rather than comprehension. Though production is not our primary interest here, we will briefly review this literature as it gives insights into children's understanding of iconicity in general.

The first gestures that children produce are deictics: points, grabs, or other gestures that indexically pick out a referent from the immediate environment (and whose meaning is therefore context-dependent). Deictics are first produced aged eight-to-twelve months (Goldin-Meadow, 2014). These early gestures actually precede the first words, with gestural reference for a given object preceding reference in speech by an average of three months (Iverson & Goldin-Meadow, 2005). They continue to lead spoken language through the one-word phase, with supplementary word-gesture combinations (e.g. *give* + deictic gesture to chocolate to indicate 'give chocolate') preceding two-word utterances by an average of 2.3 months.

---

<sup>31</sup> Unless noted, the children observed in all reported studies are (North American) Anglophones, which may have implications for the generalisability of the results. Variability in children's relationship to gesture by native language will be discussed later.

Deictic gesture use not only precedes but *predicts* rate of acquisition of spoken language. The onset of word + deictic gesture combinations predicts the onset of the two-word stage (Iverson & Goldin-Meadow, 2005), with the relationship only breaking down later when children start producing more complex constructions like ditransitives (Özçalışkan & Goldin-Meadow, 2009). Gesture also predicts vocabulary acquisition: gesture production at 14 months predicts vocabulary size at 54 months controlling for speech (Rowe & Goldin-Meadow, 2009), and gesture at 24 months predicts vocabulary at 42 months, over and above child and parent vocabulary and socio-economic status (Rowe, Özçalışkan, & Goldin-Meadow, 2008). One possible explanation for the relationship between infant gesture production and acquisition of spoken language is that caregivers frequently ‘translate’ gestural or word + gesture productions by children into well-formed speech (Goldin-Meadow, Goodrich, Sauer, & Iverson, 2007), effectively giving children the ability to solicit tuition in the constructions and vocabulary they are interested in producing.

Iconic gestures appear later than deictics. Children’s deictic gestures are largely used in place of nouns, but iconics tend to replace verbs or, less frequently, describe object attributes (Özçalışkan et al., 2014; Özçalışkan & Goldin-Meadow, 2011). While deictics precede the emergence of nouns, the first iconic gestures appear about six months *after* the first verbs, with little iconic gesture production before 26 months, in either naturalistic observational contexts or experimental settings (Behne, Carpenter, & Tomasello, 2014: an experiment with German speaking children; Özçalışkan, Gentner, & Goldin-Meadow, 2014 and Özçalışkan & Goldin-Meadow, 2011: observational studies of English speaking children). However, once iconic gestures do appear, a large proportion are used for meanings that are not yet present in the child’s spoken vocabulary (Özçalışkan et al., 2014). However,

Özçalışkan et al. (2014) report that iconic gestures are comparatively infrequent, appearing about 100 times less frequently than verbs. As a proportion of young children's gestures, iconic gestures also tend to be low frequency, representing perhaps 1-5% of the total (Özçalışkan & Goldin-Meadow, 2011).

However, native language appears to have a substantial effect on both the nature and the onset of children's iconic gestures. Özçalışkan et al. (2014) found that among iconic action gestures produced by 14-to-34-month-old Anglophones, gestures describing manner of motion were at least six times more frequent than those describing path of motion, mirroring English's preponderance of manner verbs. By contrast, Gullberg, Hendriks, & Hickmann (2008) report that Francophone children aged four to six years largely produce gesture for path when describing motion scenes, in keeping with adult patterns reflecting the dominance of path verbs in French (though see Özyürek et al., 2008, for evidence that gestural differences between English and Turkish speakers only emerge after age five). Turning to age of onset, Furman, Küntay, & Özyürek (2014), who analysed excerpts from a video corpus in which Turkish children discuss caused motion, report that iconic gestures emerge at 22.5 months on average, several months earlier than the norm for English children. Furthermore, iconic gestures were as frequent as pointing gestures (though it's worth noting that discussion of caused motion seems particularly conducive to iconic gesture). Ninety two percent of the iconic gestures encoded action, and Furman et al. attribute their early emergence to the early onset and high frequency of verbs in Turkish, arguing that iconic gestures require verb knowledge as a semantic model, but pose no special cognitive challenges beyond this. In support of the claim that children need language-specific knowledge to guide their use of iconic gesture, Nicoladis (2002) reports that free-playing French-English bilingual children aged

three-and-a-half to five use more iconic gestures in the language in which they are more proficient, even though this is not true of other gesture types.

It should be noted though that there is a lack of individual studies with controlled cross-linguistic comparisons: the experiments above differ both in terms of the age of subjects and the nature of the tasks and observations used, and should therefore be interpreted with caution.

### **Do children comprehend iconic gestures?**

The standard assumption in the literature is that the ability to comprehend iconic gestures emerges at the same time as the ability to produce them: i.e. 26 months. However there are only a handful of studies that explicitly test this hypothesis, and the results of these are mixed. Striano, Rochat, and Legerstee (2003) did find that 26- but not 20-month-olds performed above chance at choosing which of four objects to send down a slide to match an iconic gesture. However, 20-month-olds *were* able to perform above chance when the affordances of the objects were modelled beforehand, suggesting that the failure of comprehension could be as much to do with a lack of background on how objects are used as with a lack of iconicity comprehension *per se*.

Other studies suggest that mastering comprehension of iconic gesture is an incremental process, both before and after the 26-month threshold. Namy (2008), testing 14-, 18-, 22-, and 26-month-olds, presented children with a novel iconic gesture and required them to use it to choose between two objects. She found some evidence that 18- and 22-month-olds could succeed at this task in addition to 26-month-olds. Other studies suggest that understanding continues to develop after 26

months. Tolar, Lederberg, Gokhale, and Tomasello (2007) tested whether children aged between two-and-a-half and five years could successfully choose between four photographs on the basis of an iconic gesture. All age groups performed above chance, but children performed incrementally better with increasing age. Interestingly, chronological age was a better predictor of performance than language age. Stanfield, Williamson, and Özçalışkan (2014) tested two-year-olds, three-year-olds, and four-year-olds on their ability to integrate information from speech and iconic gesture. Children were presented with speech + iconic gesture combinations. In each case the gesture included disambiguating supplementary information (e.g. 'I am eating' + move cupped palms towards mouth in parallel as if eating a sandwich). Children were then shown two photos of objects (e.g. breakfast cereal vs. sandwich), only one of which was targeted by the gesture, and asked to choose the right one (e.g. 'What did I eat?'). Two-year-olds performed at chance, but three- and four-year-olds both did better than chance, with four-year-olds outperforming three-year-olds.

However, though these studies aim to test children's understanding of iconic gestures, they impose additional demands beyond comprehension, a point I will return to at the end of this introduction.

### **Iconic gestures and word learning**

If iconic gestures play a role in vocabulary acquisition then children need to not only comprehend them, but also apply the information this gleans them to the task of word learning. And indeed, recent studies have shown that young children can use iconic gestures to make inferences about the meaning of verbs. Goodrich and Kam (2009) showed two-, three-, and four-year-olds pairs of toys, each of which exhibited

a distinctive form of movement. When children were introduced to a new verb accompanied by an iconic gesture picking out one of the two toys, all age groups succeeded in picking the right toy on the basis of a repetition of the word and gesture. In a follow up study (from which two-year-olds were excluded due to task memory load), three- and four-year-olds were again shown a pair of toys, but then introduced to *two* pairs of novel verbs and iconic gestures. Both age groups were able to select the right toy on the basis of the repetition of one of the verbs alone (without supporting gesture). This means that the children had paired the novel verbs with information from the accompanying gesture, held this in memory (however briefly), and then deployed this information when they encountered the verb again.

Children can also use iconic gestures to resolve Quinean ambiguity as to which aspect of a scene a verb encodes. Mumford and Kita (2014) showed three-year olds videos of a novel action, while naming the action using a novel word accompanied by an iconic co-speech gesture. The iconic gesture represented either the manner of the action, or its end state. Children generalised the verb to other videos with the same following manner gestures, but generalised to the same end state after end-state gestures.

### **Possible limiting factors on infants' mastery of iconic gestures**

Children's apparent inability to reliably produce iconic gestures before the age of 26 months does not reflect a general lack of gesture comprehension. As children's early production of deictic gesture suggests, children as young as 13 months understand that pointing is referential (Gliga & Csibra, 2009). What hurdles do they need to cross before iconic gesture can emerge?

Özçalışkan, Genter, & Goldin-Meadow (2014) discuss this question at length, considering a range of quite distinct possibilities (which are nonetheless not necessarily mutually exclusive). They argue that action meanings are inherently difficult compared to noun meanings; actions differ from objects in that they have no basic level description, and therefore different languages focus on different aspects of action (e.g. path versus manner-of-motion; cf. Gentner, 1982). Before iconic gestures take hold, children need to learn their language's schema for encoding actions in verbs. This would explain both why iconic gestures trail the first verbs (reversing the pattern seen for deictic gestures and nouns) and emerge earlier in languages where verbs are learned earlier (Furman, Küntay, & Özyürek, 2014), and why early iconic gestures appear to mirror the properties of the verbs in the child's native language. This view parallels that expressed by Furman et al. that iconic gesture production depends crucially on the verb semantics of a child's native language (though against this Tolar, Lederberg, Gokhale, & Tomasello, 2007, report that chronological age is a better predictor of iconic gesture comprehension than language age).

Özçalışkan, et al. also consider the possibility that iconic gesture critically depends on structure mapping in the sense of Genter (2010) – i.e. the ability to align representations that share isomorphic abstract structures. This makes intuitive sense – many iconic gestures lack narrow perceptual resemblance to their objects, instead depending on schematic correspondences. In support of this view, children show a wider shift in their ability to entertain relational concepts around the time they begin to comprehend iconic gestures (Loewenstein & Gentner, 2005). For instance, only at about 26-months do children show the ability to link realistic model agents to videos showing the agent in action (Johnson, Younger, & Furrer, 2005), and 30-month-olds

succeed at using a location pointed out in a picture to find a toy hidden in a room whereas 24-month-olds do not (DeLoache & Burns, 1994). Emmorey (2014) also emphasises structure mapping as a framework for understanding the effects of iconicity, though in the context of sign language grammar and processing. She also raises the important point that even if structure mapping abilities are in place, understanding of iconicity often requires conceptual background. For instance, an iconic sign MILK, which mimics a milking action, will be lost on a child that isn't familiar with the mechanics of old-fashioned dairy. There are hints in the literature that lack of conceptual knowledge of the target of a gesture may sometimes limit young children's understanding of iconic gestures. For instance, as discussed above Striano, Rochat, & Legerstee (2003) found that 20-month-olds can comprehend iconic gestures, but only after seeing an experimenter model the object affordances the gestures target.

Özçalışkan, et al. point out that Liszkowski (2010) has suggested that iconic action gestures may be inherently more difficult than words for young children because their comprehension requires decoupling the gesture from the goal the action schema suggests, and reinterpreting it symbolically. This suggestion draws support from work by DeLoache (2002) showing that even at 30 months, children have difficulty using a scale model to find a toy hidden in a room, but that this difficulty is reduced when children are told that the model is the room after having been shrunk, thus removing the need to conceptualise it as a representation as well as an object in itself.

Finally, Özçalışkan, et al. go on to discuss the possibility that iconic gesture input from parents is sparse until 26 months. This might indeed explain the emergence of child iconic gesture at this age: Goodwyn, Acredolo, & Brown (2000) found that

instructing parents to use iconic gestures substantially increases child iconic gestural repertoire. However, this explanation raises the further question of why parental use of iconic gesture increases prior to 26 months, and whether it isn't driven by increased receptivity on the child's part as a result of other aspects of cognitive development.

In summary, several factors could contribute to the emergence of iconic gestures in the third year of life. It is plausible that multiple conditions must be met: an understanding of verb semantics, an ability to structure map, sufficient conceptual background to link iconic gestures with their targets, precedents in the input, and an ability to view action-like gestures representationally. Nonetheless, this discussion assumes that comprehension (as opposed to merely production) of iconic gesture only emerges in the third year of life, a claim that is still under-supported. I will return to this point at the end of the introduction and the rest of the chapter.

### **Are there iconic gestures in child-directed communication?**

Özçalışkan, Gentner, & Goldin-Meadow (2014) tout lack of iconic gestures in the input as a possible reason for the comparatively late emergence of iconic gesture production. As discussed, iconicity is pervasive in the co-speech gestures adults produce when talking to other adults (which it should be noted are certainly a potential part of the input). But how much iconic gesture is there in child-directed communication? Obviously iconic gesture in the input is just as much a prerequisite for it playing a role in vocabulary acquisition as children being cognitively equipped to use that gesture. The input question is under-studied compared to children's production and comprehension, but there are a handful of studies that yield useful

information. The picture they paint is of parents' iconic gestures being comparatively infrequent, but increasing in frequency as children mature.

Iverson, Capirci, Longobardi, & Caselli (1999) observed 12 Italian mother-child pairs at home when children were 16 and 20 months old. Each observation lasted for 45 minutes, and focused on play and snack/meal time. On average, mothers produced approximately three iconic gestures at the 16 month session, and five at the 20 month session. This compares to about 30-40 deictic and 30-40 conventional gestures per session.

Rowe, Özçalışkan and Goldin-Meadow (2008) report no correlation between parental gesture at 14 months, and child vocabulary at 42 months, which would appear to argue against iconic gestures being an important cue to word meaning. Özçalışkan and Goldin-Meadow (2011) tracked children and parents from 14 to 34 months of age, recording 90 minutes of video of parent-child interactions in the home once every four months. They report that both children *and* adults undergo an iconic gesture spurt when children are around the age of 26 months, perhaps reflecting parental awareness that children are ready to understand iconic gestures at this age. However iconic gestures remain a fairly small proportion of the parental total (no more than 5%). It should be noted though that parents' use of iconic gestures may depend on the situation at hand - Gogate, Bahrick, and Watson (2000) report that parents frequently use iconic gestures when asked to teach children of 30 months and under novel motion verbs.

Moreover, Perniss and Vigliocco (2014) speculate that iconicity may be underreported in the results of many studies due to coding decisions. For instance a parent picking up a toy and demonstrating an action with it would typically be coded

as deictic rather than iconic, even if the action demonstration is effectively an iconic gesture. Moreover, children may be able to use iconic gestures in adult-adult discourse observed as a third party, even if such gestures are missing from child-directed communication. This suggestion gains some support from Perniss, Lu, Morgan, and Vigliocco's (under revision) finding that deaf mothers heighten iconicity in the linguistic input to toddlers.

### **Outstanding questions in child iconic gesture**

The literature on children and iconic gesture is piecemeal: many of the informative studies are primarily interested in other things, and even those that are aimed at iconic gesture use quite varied methodologies and samples, making it difficult to triangulate the important developmental factors by comparing studies. Moreover, there is a paucity of studies that attempt to go beyond merely noting that a particular ability is in place by a particular age, and instead attempt to dissect the suite of cognitive abilities that have to be in place to support iconic gesture, when these abilities emerge, and what factors determine their emergence. The result is that much remains unclear about the development of iconic gestures and their importance to language acquisition. Though I will not be able to address all of the outstanding questions on the development of iconic gesture in the experiment that occupies the rest of this chapter, I would like to conclude this review of the literature by identifying several lines of inquiry that I believe future work needs to explore.

Firstly: what drives the spurt in iconic gesture production that (for Anglophones) takes place around 26-months? Evidence for acquisition of verb semantics playing a role seems strong, but most of the cross-linguistic comparisons haven't matched like

for like, instead comparing e.g. naturalistic observation in one language to experimental scenarios in another. Intuition suggests that an understanding of structure mapping ought to be a prerequisite for iconic gesture, but no studies I am aware of have looked for a correlation at the level of the individual child between understanding of iconic gesture and structure mapping abilities in other domains.

Secondly: can children use iconic gestures to learn about word meanings outside narrow, short-term contexts? Goodrich and Kam (2009) and Mumford and Kita (2014) present evidence that children successfully use iconic gesture to resolve lexical ambiguity. However, in each case, the gesture and the referent of its accompanying word were either available to the child at the same time or only separated by a very short interval. But are children capable of using iconic gestures to learn about word meanings in a broader range of contexts? Can children retain information for longer periods, and use it without a narrow range of possible referents having already been identified?

Thirdly: can children learn from adult-to-adult iconic gesture? Iconic gesture occurs in adult-adult communication considerably more frequently than in parent-child communication. If exposure to third party conversations is an important part of the linguistic input, then adult-adult iconic gesture may be a rich source of data on word meaning for the child. To the best of my knowledge, no study has addressed this question.

Fourthly: does it even make sense to talk about *the emergence of iconic gesture*, as if we can assume that comprehension emerges all at once? Quite plausibly, use of iconicity involves a heterogeneous set of cognitive systems (see the Chapter 1), each of which may develop at different times. If that is the case then there will be no

single point at which the ability to comprehend iconicity appears, but rather a whole set of developmental schedules for different abilities. Given this, the right answer to the question of whether children of a given age can use iconicity might well be: what kind of iconicity, deployed how?

Finally: does the comprehension of iconic gestures precede their production? Most studies on children and iconic gesture focus on production. These find that for English speaking children at least, the ability to use iconic gesture appears around 26 months. The more limited literature on comprehension of iconic gesture tends to claim that comprehension appears at the same time as production, but in fact most studies have not even tried testing children younger than 26 months. Therefore this reasoning is rather weak, especially as researchers on language acquisition have often argued that children are systematically conservative in production: omitting constructions they comprehend perfectly well until they are fully confident in their understanding of the construction's grammatical basis (e.g. Snyder, 2011). If something similar applies to iconic gestures (which, as discussed above, may depend on acquiring grammatical features of the native language: Özçalışkan, Gentner, & Goldin-Meadow, 2014), then absence of production cannot be taken to mean that children are entirely incapable of comprehension.

Moreover, there are other reasons children might comprehend but not produce iconic gestures. Perhaps the gestures are too motorically demanding, or perhaps children are reminded of the objects, actions, properties, and events that iconic gestures encode without understanding that such gestures can be used as an intentional communicative strategy. It's also possible that the capacity to comprehend iconic gesture is sometimes impeded by children's lack of experience with the properties targeted by the gestures.

Is it therefore possible that the potential for understanding of iconic gesture is present earlier than 26 months even in English-speaking children, and that this is sometimes masked by other factors? Namy (2008) presents results that suggest that Anglophone children as young as 18 months old have some comprehension of iconic gesture, and Striano, Rochat, & Legerstee (2003), found that 20-month-olds could interpret iconic gesture once the object affordance the gesture targeted had been modelled for them.

The experiment presented in this chapter represents an attempt to start answering this question: is robust comprehension of iconic gesture present before 26 months? The few previous studies to have addressed this question showed mixed results (Namy, 2008; Striano, Rochat, & Legerstee, 2003), but both required children to execute some task on the basis of the iconicity they are exposed to. Therefore successfully completing these tasks required children to do (at least) three things:

- 1) Perceive a resemblance between a gesture and its reference in the world, based on perceptual similarity or structural isomorphism.
- 2) Understand that the gesture was a form of intentional communication, supposed to direct their behaviour in a referent choosing task.
- 3) To cooperate in choosing a referent on the basis of the gesture.

Failure in any one of these three steps would result in failure to complete the task. Thus inasmuch as children under 26 months appear to show limited understanding of iconic gesture, in these kinds of tasks, it's entirely possible that say 1) is in place, whereas 2) and/or 3) are absent.

In order to attempt to pinpoint the abilities of young children with iconic gesture we ran 18-month-olds on a much simpler task than has previously been attempted, primarily aimed at testing point 1). The advantage of simplifying the task is that it should give us a much clearer idea of what children can and can't do, and a clearer sense of what their overall abilities with iconicity may be, and of what factors limit those abilities. Arguably 1) is necessary but not sufficient for comprehension of iconicity: true mastery of iconicity requires understanding both that a sign bears a resemblance to its meaning in the world, and that is representational (in the sense of having semantic content; see Chapter 1). However, establishing exactly which components of iconicity comprehension are in place is vital to the task of understanding its development, and may moreover inform our understanding of the cognitive bases of iconicity, which would in turn inform how we taxonomise iconicity (all under-researched questions).

## Experiment

We opted for the simplest kind of experiment available in developmental psychology: a preferential looking study (Fantz, 1965). The advantage of this is that it places minimal demands on children, removing factors that might impede their expression of iconic gesture comprehension even if such comprehension is present.

The basic methodology was to show 18-month-olds videos of an iconic gesture, followed by two simultaneous, side-by-side videos of objects moving, one of which corresponded to the gesture, one of which was unrelated. We were simply interested in whether children were more likely to look at one video than the other (a preference

for either the congruent or the incongruent video would demonstrate an ability to perceive iconic similarity between gesture and object motion).

## Methods

**Subjects** were 22 17-to-19-month-olds recruited from the University of Indiana Bloomington child subject database: eight in the control condition (five females,  $M = 557.5$  days  $\pm 20.7$ , all native Anglophones, though two were in Spanish immersion daycare), and 14 in the gesture condition (five females,  $M = 557.5$  days  $\pm 13.9$ , 11 native Anglophones, an English-Spanish bilingual, an English-Cantonese bilingual, and one Marathi speaker). A Bayesian t-test carried out in R's *BayesFactor* package comparing the null hypothesis that the two groups are the same age to an alternative hypothesis that they differ in age by 30 days finds a Bayes factor of greater than  $10^{20}$  in favour of the null, allowing us to conclude with great reliability that the two samples can be taken to be within a month of each other in age (Morey & Rouder, 2015).

**Materials** Comprised two sets of six videos: six *object videos*, of various novel, eye-catching objects (recycled from the shape bias studies of Cantrell & Smith, 2013; Samuelson & Smith, 2005; and other related studies) each of which moved in a distinctive manner; and six pairs of videos of a model verbally soliciting the child's attention, and then producing a gesture iconically corresponding to the motion in one of the object videos.

All videos were in colour with 640\*428 pixels.



*Figure 1: A still from one of the object videos. The arrow represents the manner of motion shown in the video.*

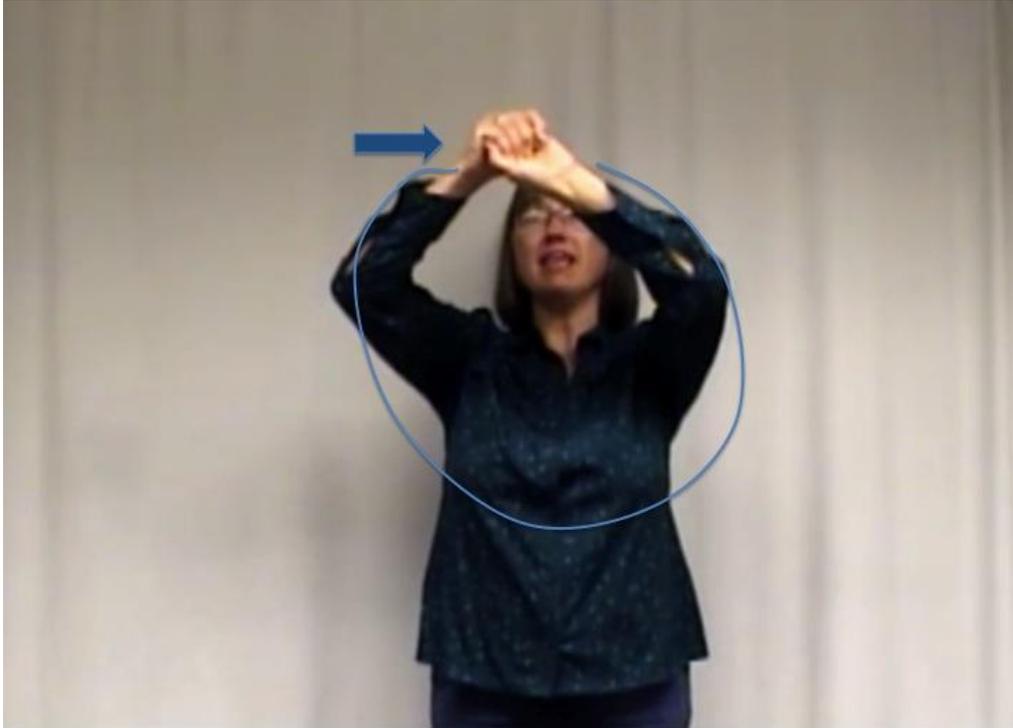
Object videos were filmed using coloured objects against a white background. The objects were chosen both so as to be visually engaging to the children, and so as to be novel. We reasoned that novelty was desirable because this way mutual exclusivity did not rule out the nonwords in the gesture video being paired with the objects in the object videos by any children inclined to interpret the nonwords as nouns (though the nonwords were intended to be taken as verbs, and given verb morphology, we reasoned that this was still a possibility). Familiar objects for which the children already knew names might have led to the nonwords (and the gestures with which they were paired) simply being ignored by such children when object videos appeared.

No human beings or human hands appeared in the videos (to avoid potential for direct matching of actions to motorically equivalent gestures by the children). The

objects were instead moved using thin strings that were subtle or completely invisible in the videos. Each video was exactly 3.5 seconds long. Motion in each video was smooth and continuous, lasting for the whole duration of the video (see Table 2 for a complete list of object videos, and Appendix 1 for screenshots from each).

<b>Motion</b>	<b>Object</b>	<b>Object colour</b>	<b>Backdrop</b>	<b>Notes</b>
Approach	Throne	Dark brown	Across surface of table	Approaches camera
Bounce	Dolmen	Emerald green	In front of curtain	Proceeds left to right in 3 bounces
Circle	Lightning bolt	Pea green, red edges	In front of curtain	Circle is clockwise
Down	Peg	Dark Purple	Across surface of table	Filmed from above (to create vertical movement)
Up	Cone	Orange	Across surface of table	Filmed from above to create vertical movement)
Withdraw	Buggy	Pine green	Across surface of table	Withdraws from camera

*Table 1: List of object videos*



*Figure 2: A still from one of the gesture videos. The blue arrow and line represent the gesture motion.*

Gesture videos were variable in length (see Table 2). Each began with a verbal attention bid in which the model named the action she was about to gesture using a nonword, e.g. “Look baby, gibing!”. The gesture followed. The model stood in front of the camera, looking into it when addressing the child, and following her hands with her eyes when producing the gesture. She began each video with her hands at rest by her side, and kept them there during the attention bid. She stood far enough from the camera that her hands were in shot when at rest by her side, and when raised above her head (see Figure 2, and Appendix 2 for screenshots of all gesture videos). There were two videos for each gesture, differing in the nonword paired with the action (so as to control for any possible sound-based iconicity in the names). These were counterbalanced between subjects.

Video	Duration (/seconds)	Nonword	Attention bid
Approach A	7	[gɑɪbɪŋ]	"Look baby, gibing!"
Approach B	7	[fɪmɪŋ]	"Look, see this? Fimming!"
Bounce A	8	[pɛɪŋ]	"Look at this: peshing"
Bounce B	7	[dʒəʊpɪŋ]	"Uh oh, look! Joping!"
Circle A	9	[fɪmɪŋ]	"Look, see this? Fimming!"
Circle B	9	[jɔfɪŋ]	"Look, see this? Yorfing!"
Down A	10	[wæzɪŋ]	"Look, see this? Wazzing!"
Down B	10	[pɛɪŋ]	"Look, see this? Peshing!"
Up A	6	[jɔfɪŋ]	"Look baby, yorfing!"
Up B	8	[wæzɪŋ]	"Look baby, wazzing!"
Withdraw A	7	[dʒəʊpɪŋ]	"Look, see this? Joping!"
Withdraw B	8	[gɑɪbɪŋ]	"Look, see this? Gibing!"

*Table 2: List of gesture videos*

## **Procedure**

Subjects arrived at the lab with parents or caregivers and were greeted by an experimenter. Before testing, the subjects were taken to a playroom equipped with various toys to acclimatise them to the lab while their caregiver filled in the Macarthur Communicative Development Inventory (standard lab procedure), and the consent form. Subjects remained in the toy room for 5-15 minutes before testing (as guided by parental judgment of when the subject was ready).

Once ready, the subject, caregiver, and experimenter moved to a small testing room. The caregiver sat on a standard kitchen-style chair positioned with its front legs one metre in front of a projector screen of approximate 1\*1.3m. The subject sat on the caregiver's lap. The door was closed and the lights turned off. Walls were covered by

white curtains, behind which the experimenter retreated to set the slide show playing, leaving nothing to distract subjects from the screen. While the children watched the two-minute slideshow, they were filmed by a camera hidden under the projector screen (frame rate: one frame per 41ms), so that their gaze direction could be coded later. Caregivers were asked to keep their eyes closed throughout the slideshow so as not to bias the subject's gaze.

Once the slideshow was over, both caregivers and subjects were thanked, and the subject was given a small gift (e.g. a book, ball, or bag).

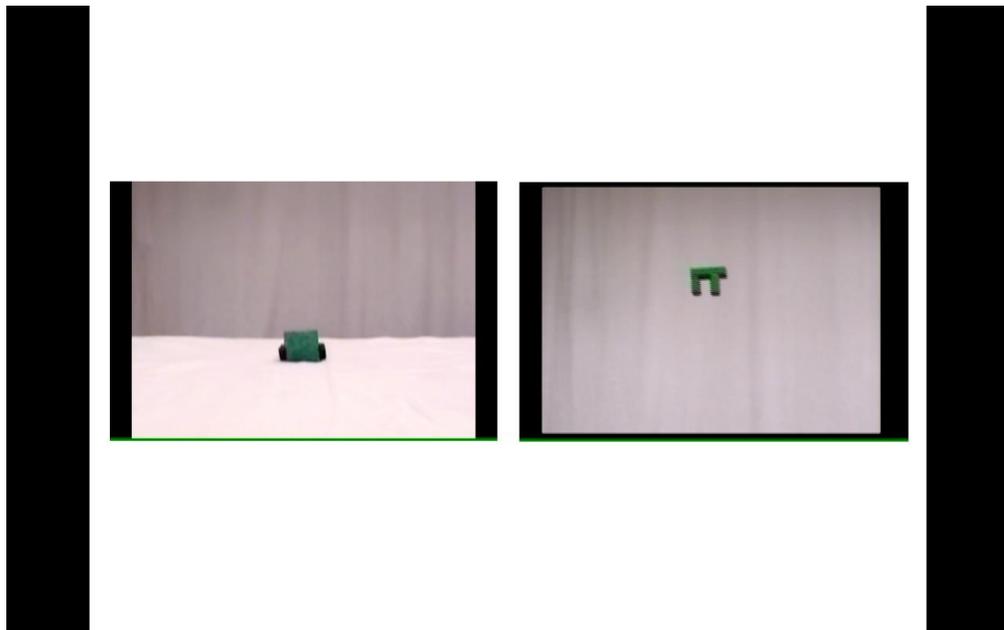
Subjects could either be in the gesture condition, or the control condition. The control condition was identical to the gesture condition, except that the gesture video was replaced by an unrelated attention-getter. The purpose of the control condition was to control for the possibility that, in the pairings we used, certain target object videos were inherently more interesting to the children than their foils (or vice versa), independently of the iconic mapping from gesture to object motion. **Note that in the control condition only phase 2) (below) was changed. All other phases, including the crucial data-providing phase 3) (with its particular nonword), remained identical**, with the control videos later being coded in the same way as the gesture videos (i.e. treating the same object video as the target).

Each subject watched a two-minute slide show containing six trials. Each trial had the following structure:

- 1) Attention Getter (expanding Sesame Street character, duration 3s).

- 2) Gesture video. In the control condition this was replaced by a second silent animated attention getter of equal length to the gesture (a spinning Sesame Street character). Duration: 6-10s (see Table 2).
- 3) A black screen with a speaker icon in the centre. The model announces the target video using the same nonword as in phase 2) of the gesture condition, using the formula: “Can you see [e.g.] fimming?”. Duration: 3s.
- 4) Two object videos play side by side. One (the target) corresponded to the gesture, the other (the foil) was unrelated (Figure 3, see Appendix 3 for slideshow orders). This is the crucial phase of the trial from which our preferential looking data are drawn. Duration: 3.5s.
- 5) Black screen to act as a buffer before the next trial. Duration: 1s.

Foils were drawn from among the other five object videos.

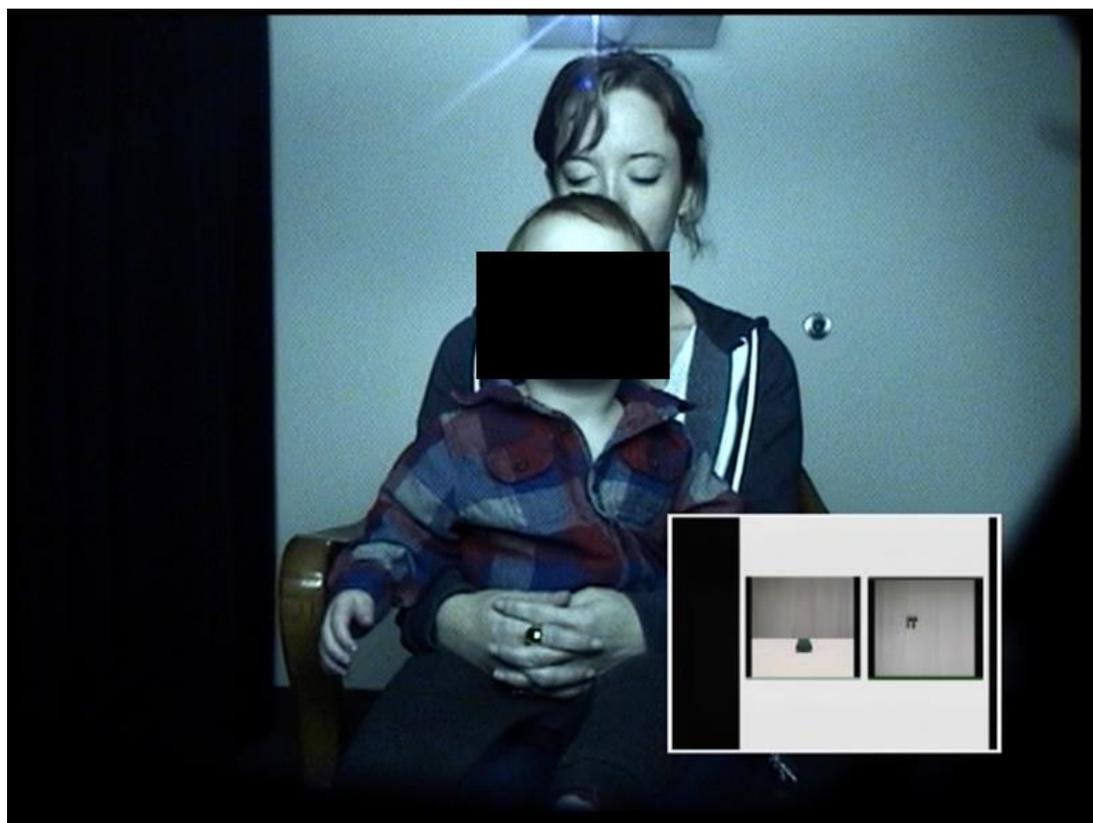


*Figure 3: A still from Stage 4) of a trial, with the target on the right, and the foil on the left. This is the stage of the trial from which our data are drawn: here a look to the*

*right of the screen by the child would have been coded as a look to target, a look to the left as a look to foil. The foil will appear as the target video in a trial later on.*

Trial order and pairing of gestures with words was randomised across four separate slideshows (see Appendix 4 for specifications), which served to counterbalance trial order, target-foil pairing, and gesture-nonword pairing. Two children were run on each of the four control slideshows, four were run on gesture slideshows 1 and 4, three on gesture slideshows two and three.

## Results



*Figure 4: A screenshot from the video capture of a subject. The image projected in front of the subject appears in the bottom right of the screen, to allow syncing of*

*gaze to slideshow. The subject's gaze can be readily coded: here he is looking at the target (right).*

**Coding** was conducted manually by experimenters, who coded the object video phase, and were blind to the content of the preceding gesture video phase. Each frame of the subject capture videos was coded according to whether the child was looking at the left of the screen, the right of the screen, or elsewhere (meaning that the child was looking away from the screen, or that their eyes were closed). Frames occurred at 41ms intervals.

Temporally, the data used in the analysis were arranged by frame number of the object video phase (Phase 4 above), coded from first look to the screen. By default, Frame 1 was the first frame of the phase. However, *on trials where subjects were not looking at the screen at the beginning of the phase, Frame 1 was taken to be the first frame in which they did look at the screen.* Any looks away from the screen following the first look did not interrupt counting of frame number. Thus if Phase 4) is 90 frames long, then a trial on which the subject is looking at the screen on the first frame but looks away from Frame 41 to Frame 60 would result in 90 frames of data, whereas a trial on which the subject starts the phase looking away from the screen and only looks at the screen for the first time on the 21<sup>st</sup> frame would result in 70 frames of data, with Frame 1 corresponding to the frame on which the first look occurred.

**Exclusion criteria** Trials in which subjects attended to the screen for fewer than 75% of frames in Phase 2) (the gesture phase) were excluded on the basis that if the child had not been paying attention to the gesture, looks to the target/foil would not be informative. Even though Phase 2) of the control condition only features an

attention getter, the same criterion was applied in order to avoid e.g. systematically excluding results from unengaged subjects from the gesture condition but not the control condition, and thereby confounding results. In total, three trials from three different subjects were excluded from the gesture condition, and nine trials from four separate subjects were excluded from the control condition.

**Analyses** used linear mixed effects models executed in the lme4 package for R (Bates, Maechler, Bolker, & Walker, 2015).

The first analysis was a simple binomial model testing the null hypothesis that subjects were no more likely to make their first look to the target than the foil (or vice versa). The model featured the single predictor of condition, random intercepts by participants, and random intercepts and slopes (for condition) by target video. Condition was coded as gesture condition = 0.5, control condition = -0.5. However, neither the intercept<sup>32</sup> ( $\beta = 0.343$ , 95% CI [-0.305, 1.029],  $z = 1.224$ ,  $p = .221$ ), nor the condition predictor ( $\beta = 0.690$ , 95% CI [-1.852, 0.362],  $z = -1.503$ ,  $p = .133$ ) were significantly different from zero, indicating that we cannot reject the hypothesis that subjects are equally likely to take a first look at either target or foil in both conditions.

The next analysis was of length of individual looks to target versus foil. The length of each continuous look to either target or foil during the object video phase was calculated. Visual analysis of a histogram of the data suggested that they would be closer to normality log-transformed (the natural logarithm was used). Once log-transformed, three visually identified outliers (all looks of one or two frames in length) were removed from the data. The model had two predictors: condition (gesture: 0.5,

---

<sup>32</sup> Confidence intervals here are estimated using R's *confint* function.

vs. control: -0.5), and direction of look (target: 0.5, vs. foil: -0.5). It had random intercepts for both subjects and target items, and random slopes for direction of look by both subjects and target items. An effect of gesture here would be expected to show up as an interaction between condition and direction of look: this would indicate a greater tendency to look longer at the target in one or other condition.

However, neither the control predictor ( $\beta = -0.017$ , 95% CI [-0.182, 0.148],  $t = -0.202$ ,  $p = 0.842$ )<sup>33</sup>, nor the target vs. foil predictor ( $\beta = 0.059$ , 95% CI [-0.211, 0.328],  $t = 0.426$ ,  $p = 0.686$ ), nor the interaction between the two ( $\beta = -0.021$ , 95% CI [-0.335, 0.293],  $t = -0.132$ ,  $p = 0.896$ ) were near significance.

Following up these overall analyses, I now move on to analyses that deal with the time course of the trials frame-by-frame.

### **Analysis by frame**

An omnibus binomial model was constructed to test whether condition type had any influence on participants' looking patterns over the course of trials. The dependent variable was binary: whether the participant was looking at the target video, or the foil video (frames where participants were looking at neither were excluded). The predictors were condition (gesture versus control, coded 0.5 and -0.5 respectively), and frame number – both linearly and quadratically (the latter in order to capture changes in strength of the linear effect of the time course of a trial; raw square values were used). All interactions (i.e. condition x linear generation and condition x quadratic generation) were also included. The maximum length of a trial was 86

---

<sup>33</sup> Confidence intervals here are based on the parameters' standard error in the model, as R's *confint* function failed to converge on a set of answers.

frames. Frame number was centred and scaled such that frame 1 was -1, frame 86 was 1, and intermediate frames fell between those values). As in previous chapters, I opted for a maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013). Random intercepts were included for participant and target video, random slopes were included for condition (by target video) and by frame number, both linear and quadratic (by both participant and target video).

No parameters in the model turned out to be significant ( $|z| < 0.8$  and  $p > .4$  in every case), suggesting that we have no evidence that looking behaviour is not random, and meaning there is no call for further analysis. This was the case for the intercept (which effectively captures the overall probability of looking at the target versus the foil;  $\beta = 0.111$ , 95% CI [-0.413, 0.636],  $z = 0.417$ ,  $p = 0.677$ ); linear frame number ( $\beta = -0.364$ , 95% CI [-1.632, 0.904],  $z = -0.562$ ,  $p = 0.574$ ); quadratic frame number ( $\beta = 0.473$ , 95% CI [-0.832, 1.777],  $z = 0.710$ ,  $p = 0.478$ ); condition ( $\beta = -0.136$ , 95% CI [-0.968, 0.697],  $z = -0.320$ ,  $p = 0.749$ ); the interaction between condition and linear frame number ( $\beta = 0.287$ , 95% CI [-1.424, 1.998],  $z = 0.329$ ,  $p = 0.742$ ); and the interactions between condition and quadratic frame number ( $\beta = -0.667$ , 95% CI [-4.065, 2.732],  $z = -0.385$ ,  $p = 0.701$ ). Nonetheless, I will provide a visual summary of the data, the construction of which I will explain below.

The basic approach was to split up the data by frame number and run a separate model for each frame. The graphs display parameters from frame-by-frame models of looking behaviour. I present three graphs: one for the gesture condition, one for the control condition, and one for the comparison (i.e. the difference) between those conditions.

Frame numbers were capped at 83 (this corresponds to the length of the shorter trials: there was a small amount of variation in number of frames per trial). The first 5-10 frames of phase 4) were often hard to code because the screen lighting up again after phase 3) caused a momentary bleaching out of the video recording capturing the subject. This means that these frame numbers are comparatively low power (where ambiguous, I assumed that the first frame of the phase should count as Frame 1, but I coded the subject as looking at neither target nor foil). This is reflected in wider error bars for earlier frames in some graphs.

Analysis is by mixed effects binomial logistic regression models featuring random intercepts for participants, and random intercepts and slopes for target items (the latter only in the case of comparisons between conditions), from which models were extracted a parameter estimate and 95% confidence intervals (derived using R's *confint* function). The dependent variable was the subjects' direction of gaze during a given frame in phase 4): to the target video or to the foil video. These were then plotted in a single graph. Therefore for the condition-specific graphs the dependent variable is *probability* of looking at the target rather than the foil during a given frame. For the comparison between the conditions I plot the model parameter for the difference between the two conditions. This corresponds to the models' estimate for:

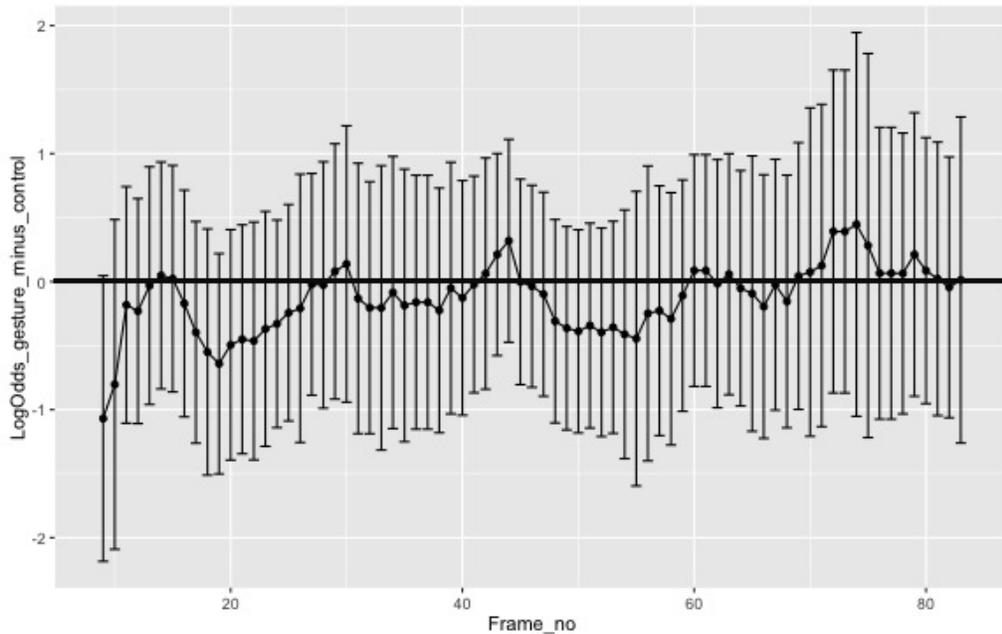
$$\ln(P(\text{look at target in gesture condition})/P(\text{look at foil in gesture condition})) - \ln(P(\text{look at target in control condition})/P(\text{look at foil in control condition}))$$

This isn't interpretable as a raw probability unless you make additional assumptions about the overall probability of looking at the target across conditions, so I plot it as it is. However, a parameter of > 0 implies that subjects are looking at the target more

in the gesture than in the control condition, whereas a parameter of  $< 0$  implies the opposite.

As I used logistic regression for these graphs I excluded the third class of datapoints, where the subject is looking away from the screen: the comparison is only between frames where the subject is looking at the target and frames where the subject is looking at the foil.

See Appendix 4 for equivalent analyses by item.



*Figure 5: Frame-by-frame model parameters for the probability of looking at the target vs. the foil in the **gesture condition vs. the control condition**. The zero line (highlighted) indicates that subjects are equally likely to look at the target in either condition. Above that line subjects are more likely to look at the target in the gesture condition, and below it they are more likely to look at the target in the control condition. Error bars represents 95% confidence intervals.*

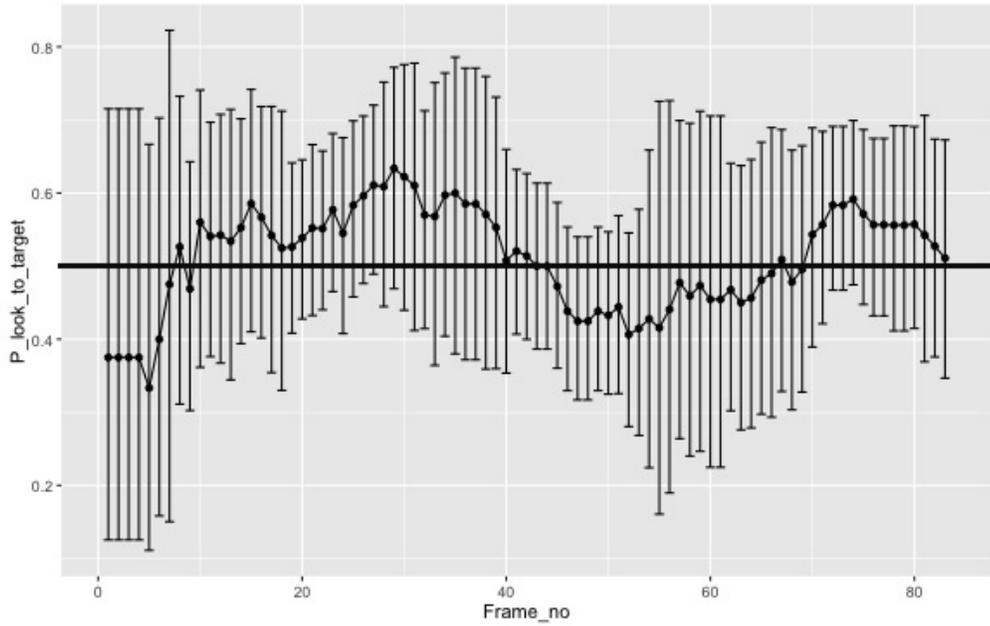


Figure 6: Frame-by-frame model predictions for the likelihood of looking to the target vs. foil in the **gesture** condition. The  $p = 0.5$  line (highlighted) indicates equal probability of looking to the target or the foil. Above this line subjects are more likely to look at the target, below it they are more likely to look at the foil. Error bars represent 95% confidence intervals.

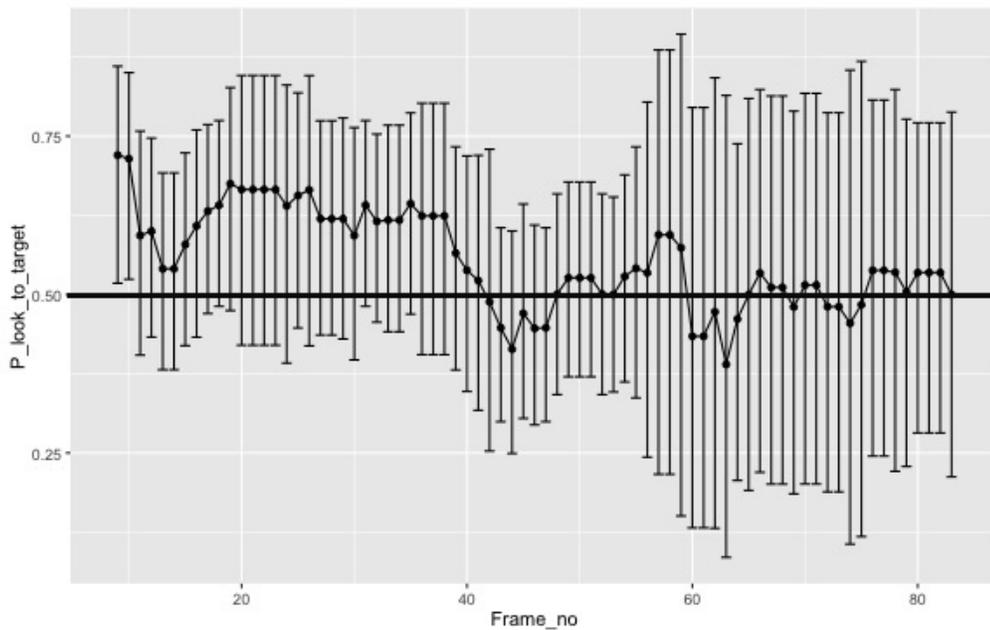


Figure 7: Frame-by-frame model predictions for the likelihood of looking to the target vs. foil in the **control** condition. The  $p = 0.5$  line (highlighted) indicates equal

*probability of looking to the target or the foil. Above this line subjects are more likely to look at the target, below it they are more likely to look at the foil. Error bars represent 95% confidence intervals.*

As can be seen, at no almost point do the lines of any of the graphs significantly differ from the origin, even without correcting for multiple comparisons, reinforcing our conclusion that we have no evidence for non-random looking behaviour. The line for the graph comparing conditions does not significantly differ from the origin at all.

## **Discussion**

To summarise: I conducted a study where children were shown videos of gestures followed by two side-by-side videos of objects moving in distinctive ways, one of which iconically corresponded to the gesture, and one of which was unrelated. In a control condition the gesture video was replaced with an uninformative attention-getter, with results coded by target and foil in the same way as in the gesture condition. We were interested in which of the object videos children tended to look at.

Our data however give us no evidence for non-random looking behaviour. This may of course reflect low power: our control condition had only eight participants. Alternatively our null result could imply that our children did not understand that the iconic gesture could be used intentionally by the model to illustrate the concept she wished to name in the gesture videos, instead regarding it as some kind of non sequitur. Alternatively, it could imply that children have difficulty dealing with action as opposed to object names in this manner: cross-linguistically, verb acquisition is

more challenging than noun acquisition (Gentner, 1982; Waxman, Fu, Arunachalam, Leddon, Geraghty, & Song, 2013). In a recent study by Arunachalam and Waxman (2011), American 24-month-olds had difficulty applying the meaning of a newly taught verb to point out the correct member of a pair of videos unless the teaching had been accompanied by comparatively rich sentential context (e.g. “The man is pilking the balloon” rather than “He’s pilking it”), which suggests that 18-month-olds may well have had trouble learning about the meaning of the comparatively bare verbs and gestures in our paradigm.

However, even if the problem is not low power, this does not necessarily imply that 18-month-old children would be unable to utilise iconicity given more exposure, or a more naturalistic setting (for instance, Sekine, Sowden, & Kita, 2015, present evidence that three-year-old children can integrate speech with iconic gesture when the two are performed live by an experimenter, but not when played from a video).

Definitive conclusions will however have to wait until more data can be collected. The data presented here were gathered during a three-month study visit to the US, and unfortunately my schedule and the limitations on recruiting children of the appropriate age restricted the size of my dataset.

Moreover, at least two aspects of the experiment’s design could be responsible for children’s apparent failure to interpret gestures as communicative, and would be worth varying in future studies. Firstly, speech and gesture were not temporally integrated, a highly artificial state of affairs that might make it impossible for children interpret the gestures in this study as non-communicative, even if they have no trouble with interpreting iconic gestures as communicative when they are integrated with speech. If this were the case then we would have to suspend all judgment about

children's ability to interpret the referentiality of iconic gestures until further work is possible. Secondly, in the crucial object video condition, both videos were novel (at least within the trial). Thus visual attention was divided, and children could not have known which video to attend to first to find a match to the gesture. This could have added noise to our data, thus in future it would be worth introducing the object videos at the start of the trial, showing the gesture, and only then showing the object videos again.

## Chapter 6: Conclusions and General Discussion

This thesis has been an exploration of the place of iconicity in spoken language. It has covered spoken and gestural iconicity, the mechanisms of iconicity, and the role of iconicity in word learning, language change, language origins, and language acquisition.

I began the introduction with a very brief genealogy of the concept of iconicity, and unpacked what I would mean by that term in the remainder of the thesis. I concluded that the proper definition of iconicity for my purposes was not some metaphysical relation between the form of a sign and its meaning, but rather a psychological relationship between how those two things are represented. I emphasised that my claim is that iconicity is likely to involve a heterogeneous set of cognitive resources, an understanding of which is a crucial task for iconicity research. I pointed out the need for findings about the cognitive basis of iconicity to inform refinement of our theoretical taxonomies and definitions of the phenomenon.

I followed this theoretical preamble with a survey of the literature on iconicity. I showed that from a position of comparative neglect during the twentieth century, which was dominated by structuralist and generativist approaches to language that had little interest in iconicity, the topic has generated a significant amount of work among psychologists and neuroscientists in the 21<sup>st</sup> Century. I presented evidence that iconicity is present in the lexica of both spoken and signed languages, that people are sensitive to iconicity, and that iconicity benefits both processing and acquisition. I sketched a possible taxonomy for different kinds of iconicity, but point out however that much remains unresolved as to its cognitive mechanisms, the

reasons it is ubiquitous in vocabulary, and its importance to acquisition – questions I go some way to addressing in the remainder of the thesis.

Chapter 2 introduced sound-shape iconicity (the “kiki-bouba” effect) in detail. I reviewed the fairly extensive literature on the effect, demonstrating that it has been shown in a wide-range of paradigms and populations, though – as Cuskley, Simner, and Kirby (2015) point out – not quite as wide a range of populations as is sometimes assumed. I discussed a number of accounts of the effect, based on lip shape, orthographic influence, brain anatomy, and learned associations between certain acoustic properties and object properties. I also gave a brief introduction to the phonetics used in the rest of the chapter. I then presented an extensive norming study of consonant-vowel syllables drawn from the phonemic inventory of English. This was a wider survey of the phonetic space than has been attempted before. My results confirm that most standard generalizations about which vowel types and consonant manners are associated with roundness vs. spikiness hold up, laying the groundwork for Chapters 3 and 4, both of which make use of sound-shape iconicity. However, simple generalisations across consonants seem difficult: for instance, roundness correlates with both sonority and voicing, but voiced fricatives like [z] are spikier than voiceless equivalents like [s]. I also checked whether my data were consistent with recent claims in the literature. Like Fort, Martin, and Peperkamp (2015) I found that consonants exert more effect on iconicity than vowels. Like Cuskley et al. I found a correlation between spikiness of consonant and spikiness of the letterform typically used to represent that consonant. However, this correlation was not perfect, and may be explicable in other terms. Following this I asked a question of my data that has been little addressed in the literature so far, which was the extent to which the place and manner of consonants, and the vowels with which

those consonants are paired, interact in determining the iconic properties of syllables (rather than just having independent additive effects). Though the questions I could ask were somewhat limited by the English consonant set, I found that the lips seems special, with bilabial consonants not only being rounder than those at other places of articulation, but also differing from each other less in their roundness. Clearly this is an interesting result from the point of view of the lip shape hypothesis of sound-shape iconicity, but I also point out ways in which other accounts might explain it. Finally, I review how well existing accounts of the effect explain my data, and conclude that none is adequate to explain it all, suggesting that multiple factors are in play in creating sound-shape iconicity.

Chapter 3 further explores sound-shape iconicity, this time using the cross-situational learning paradigm (Yu & Smith, 2007). I begin by introducing the paradigm, which is a model of word learning under conditions of referential ambiguity; just the situation we might expect iconicity to help solve some version of the *gavagai* problem. Participants are exposed to a series of trials where a word (or series of words) is presented with several potential referents. No single trial suffices to disambiguate the meaning of words, but co-occurrence patterns across trials do. The first experiment I present is a near replication of Monaghan, Mattock, and Walker (2012), who taught participants iconically congruent or incongruent names for round and spiky shapes. Like Monaghan et al., I found that iconically congruent pairings improves performance in cross-situational learning, and that this advantage is category level: i.e. only appears in trials where a congruently named target is paired with a foil shape from the opposite category (e.g. round target, spiky foil). However, while Monaghan et al. found no advantage of iconicity in the first block, and therefore concluded that iconic names helped learning in some general sense,

we found that iconicity was beneficial immediately, and that therefore iconicity was simply used to disambiguate reference (as is consistent with the between category nature of the effect). This represents an important clarification of how iconicity can aid word learning. I followed this up with three experiments of near-identical design (though slightly different verbal stimuli) designed to test whether sound-shape iconicity is stronger for one kind of pairing than the other (as might be predicted by e.g. the lip shape account), an important parameter of the effect for identifying its mechanism, but one that has barely been explicitly investigated. The concept was to have one condition where round words were pitted against iconically neutral words, and another where spiky words were pitted against neutral words. If, in the absence of the symmetry enforced by the task demands and binary forced choice nature of previous studies, one pairing is stronger, then we ought to see a stronger effect of sound-shape iconicity in that condition. The first experiment did indeed show an effect of iconicity in the round but not the spiky condition, consistent with the effect being based on rounded lip shape. However, this asymmetry failed to replicate in the next two experiments. Bayesian statistics were introduced to the reader, and then used to pool the data and look at the overall picture. They suggested that any asymmetry between round and spiky is small if it exists at all, implying that lip shape is probably not the only contributor to sound-shape iconicity.

Chapter 4 continues the investigation of spoken iconicity, this time using the iterated learning paradigm (Kirby, Cornish, & Smith, 2008), a model of cultural evolution where participants learn from their predecessors, like in the game Chinese whispers/broken telephone. I introduce the paradigm and its theoretical motivations, showing how it demonstrates how defeasible psychological biases can be amplified over generations of cultural transmission into ubiquitous features of language, which

may be the right way of looking at how iconicity gets into natural language. I show that starting from an iconically arbitrary language of names for videos of moving shapes, iconicity emerges for both shape and motion over ten generations of iterated learning. This holds whether the medium of the language is text or speech. Probing what dynamic brings about iconicity in these experiments, I carried out a follow up where subjects simply generated words for stimuli (again differing on shape and motion). These names tended to incorporate the same kinds of iconicity we saw in the iterated learning experiment, suggesting that the emergence of iconicity there was at least partly attributable to production biases, as well as superior learnability of iconic names generated by essentially random drift. I conclude that this set of experiments provides evidence that 1) biases in human cognition expressed in language evolution will create pressure towards iconic vocabulary (a pressure that is of course balanced by others – e.g. discriminability – meaning that iconicity will not accrue forever), a novel approach in the iconicity literature; and 2) that spoken segmental iconicity offers richer means for bootstrapping communication than often assumed.

Chapter 5 shifts focus from speech to gesture and from adults to toddlers. Given the relative paucity of iconicity in the lexicon of a spoke language like English, one important way it may contribute to vocabulary acquisition is through gesture (though see Laing, 2014, for evidence that onomatopoeia may be important in the very early stages of vocabulary acquisition in German). However, though Anglophone children's production of iconic gesture is sparse before 26 months, almost no studies have tested young children's comprehension, and those that have gave mixed results. I tested whether 18-month-olds were able to map iconic gestures to manners of motion using a preferential looking paradigm that removed the additional task

demands present in previous studies. My results were not significant, but this may be due to limited power. Further work will help elucidate precisely what aspects of iconicity young language acquirers master when.

## Conclusions and Future Directions

The results presented here build on the literature reviewed throughout to make a strong case for the importance of iconicity in spoken language, specifically language acquisition, processing, and change. The work in this thesis makes several new contributions. Firstly, it establishes more clearly than has yet been achieved the phonetic parameters of sound-shape iconicity (for English speakers at least), which has crucial implications for the basis of the effect (essentially ruling out any neat explanation in terms of any of the single mechanisms yet proposed). It shows that such iconicity can be useful to learners in lab models of vocabulary learning (i.e. cross-situational learning), and moreover why (i.e. referential disambiguation). It shows that spoken iconicity (namely sound-shape iconicity and motion iconicity) emerge in models of language change. This suggests that psychological biases towards iconicity can be expected to lead to it consistently emerging through the cultural evolutionary dynamics of language, potentially placing it on an equal footing with other language 'universals' (Kirby, Smith, & Brighton, 2004). Finally, I show that children younger than previously assumed have some grasp of iconic gesture, but that this grasp may be incomplete as it misses an appreciation of iconicity being a deliberate communicative strategy (though this is debatable).

More speculatively, the work I have presented could be taken as support for the claim that iconicity may be crucial to spoken language in another respect. Many have

suggested that any proto-language spoken by our ancestors at the origins of language would have begun to bootstrap communication by incorporating extensive iconicity. This iconicity would have built early language on the foundations of (putatively) prelinguistic abilities like analogical mapping. Many have argued that this must have come about through gesture due to the limited iconic potential of speech, with language only making the jump to the vocal modality later in order to free the hands (e.g., Arbib et al., 2008; Armstrong & Wilcox, 2007; Corballis, 2009; Sterelny, 2012; Tomasello, 2008). However, the results presented here, particularly those in Chapter 4, suggest that contrary to what many have assumed, vocal iconicity can carry crucial properties of objects and events (as suggested by Imai & Kita, 2014; Perniss & Vigliocco, 2014), making the gesture-first hypothesis redundant.

One of the most important theoretical conclusions of this thesis is that while non-arbitrary forms in language are much more important than once supposed, their use is not cognitively homogeneous. Any given act of understanding and producing iconicity will likely involve multiple cognitive systems, and different kinds of iconicity will be supported by different cognitive systems (compare the spontaneous structure mapping employed in interpreting novel iconic gestures to the learned sensory associations that probably at least partly underlie sound-shape iconicity). Much further work is needed to clarify this complex situation. I would suggest that the following questions are crucial:

- 1) What are the cognitive systems underlying iconicity? Specifically: How many are involved for each type of iconicity? Is there any sense in which the deployment of all forms of iconicity rely on the same resources for

establishing resemblance, or – more plausibly – does this vary completely, depending on modality and abstraction? How dissociable is resemblance mapping from interpreting this resemblance as a deliberate communicative/referential strategy?

- 2) Which of the systems underlying iconicity appear at what point during development, and therefore at what age can children start to make use of iconicity?
- 3) How should our theoretical conception of and classificatory system for iconicity be built or modified given the answers to 1) and 2)?

There are also a number of more specific future directions that it will be important to pursue in this line of research. Regarding the work on the mechanism of sound-symbolism, a crucial task will be to validate results cross-linguistically. In particular it would be instructive to rerun the experiment presented in Chapter 2 on speakers of different languages, and also on children and non-literate people - cf. Cuskley, Simner, and Kirby's (2015) suggestion that the effect is driven by orthography. Any differences between different populations will help establish how much the boundaries of the effect depend on orthography and native language (especially regarding the status of phonotactically valid vs. invalid words), and how much the effect is language-general.

Most of the studies in this thesis have been used to argue for iconicity playing a role in vocabulary acquisition and/or language change. However only one chapter presents data from children. As both of these processes primarily concern children, it is vital to replicate the kinds of experiments presented here on that population. The

conclusions drawn here will only hold fast if similar results are obtained for children. Testing children using precisely the same techniques as presented in Chapters 2-4 would doubtless present methodological problems, particularly with iterated learning, which requires participants to produce names on demand in a way that young children are likely to struggle with. However, more implicit methods can be used to test children's learning, and perhaps by testing names around the phonological space of the right one we can also test the kinds of mistakes children would likely make in production.

There is potentially a lot to be gained by simply exploring different starting conditions for the iterated learning experiments presented in Chapter 4, both in terms of names and referents. For instance: how does iconicity trade off against compositionality (which can also ease the task of word learning)? Is iconicity perhaps of more use when naming referents in a domain that is not obviously semantically compositional? This could be investigated by altering our referent set so as to more resemble the holistic, non-compositional stimuli in Verhoef, Kirby, and de Boer (2016). It may also be interesting to see what happens when the initial Generation 0 language contains an overabundance of iconicity. The results of the iterated learning experiments here suggest that the oft-repeated observation that nascent sign language lose iconicity (Frishberg, 1975) represents not a terminal decline in iconicity, but rather convergence to some non-zero equilibrium level. By starting an experiment with an excess of iconicity, we could test this claim by seeing whether this iconicity got thinned out and stabilised at the same level as we see in experiments that start with an arbitrary initial language.

Sign languages are much richer in iconicity than spoken languages. This finds a highly plausible explanation in the observation that many fewer of the concepts we

express with words map on to the sounds available in the spoken modality than contain features relating to spatial forms and paths (at least via analogy, metaphor, or metonymy). However, a question that seems to be very little addressed is why different spoken languages appear to have different amounts (or at the very least different kinds) of iconicity, the obvious distinction being between sound-symbolic languages like Japanese and languages like English that have no sound-symbolic vocabulary. I would speculate that in part this may relate to phonotactics. For example, sound-symbolic languages frequently make use of reduplication: perhaps the cultural evolution of iconicity in such languages interacts with the appearance of grammatical mechanisms (or removal of grammatical constraints) licensing such reduplication. A survey of the phonological properties of sound-symbolic vs. non-sound-symbolic languages would help establish whether this hypothesis has any basis. Beyond phonotactics as conventionally understood, it may be that there is something like a high level parameter determining whether a language allows sound-symbolism. It would thus be interesting to run an experiment that attempts to teach participants not just first order lexical information (i.e. vocabulary), but also, implicitly, second order lexical information (i.e. parameters of the vocabulary generally), and to see how this affects the course of iterated learning. This could be achieved by a prior training phase undertaken before iterated learning where participants are taught words from the same 'alien language' they will be working with in the iterated learning. In one condition the language would have semantically systematic sound-symbolism. In the other the iconic coverage would only be patchy (overall frequency of iconicity could be matched: the crucial thing is that it would be a systematic phenomenon only in one condition). The iterated learning experiment would use fresh referents (and novel, non-iconic words for the initial generation). The interesting question would then be whether sound-symbolism-like iconicity would

emerge more consistently in the first condition, where participants had been led to believe that high level parameters of the language licensed it.

Finally, I would be glad to be able to carry out a new and adequately powered version of the experiment presented in Chapter 5 in order to confirm whether or not language-acquiring toddlers are able to take meaning from iconic action gestures.

Attending to these questions, scholars of iconicity will be able to establish a richer and deeper understanding of what iconicity is, and what role it plays in language and its processing and acquisition. I look forward to this understanding taking shape.



## References

- Ahlner, F., & Zlatev, J. (2010). Cross-modal iconicity: A cognitive semiotic approach to sound symbolism. *Sign Systems Studies*, 38(1), 298-348.
- Alpher, B. (2001). Ideophones in interaction with intonation and the expression of new information in some indigenous languages of Australia. In F. K. E. Voeltz and C. Kilian-Hatz (Eds.), *Ideophones* (pp. 9-24). Amsterdam: John Benjamins.
- Arbib, M. A., Liebal, K., Pika, S. (2008). Primate vocalization, gesture, and the evolution of human language. *Current Anthropology*, 49(6):1053-76.
- Armstrong, D. F., & Wilcox, S. (2007). *The gestural origin of language*. Oxford: Oxford University Press.
- Arunachalam, S., & Waxman S. R. (2011). Grammatical form and semantic context in verb learning. *Language Learning and Development*, 7(1), 169-184.
- Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex*, 63, 196-205.
- Atoda, T., and Hoshino, K. (1995). *Giongo Gitaigo Tsukaikata Jiten [Usage Dictionary of Sound/Manner Mimetics]*. Tokyo: Sotakusha.
- Aveyard, M. (2012). Some consonants sound curvy: Effects of sound symbolism on object recognition. *Memory and Cognition*, 40(1), 83-92.
- Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, 29, 832-843.

- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278.
- Barrett, H.C., Bolyanatz, A., Crittenden, A.N., Fessler, D.M.T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., Scelza, B.A., Stich, S., von Rueden, C., Zhao, W., & Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, 1522070113v1-201522070.
- Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.
- Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: position paper. *Language Learning*, 59: Supplement 1, 1-26.
- Behne, T., Carpenter, M., & Tomasello, M. (2014). Young children create iconic gestures to inform others. *Developmental Psychology*, 50(8), 2049-60.
- Berlin, B. (1994). Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature. In L. Hinton, J. Nichols, and J. J. Ohala (Eds.), *Sound Symbolism* (pp. 76-103). Cambridge: Cambridge University Press.
- Berlin, B. (2006). The first congress of ethnozoological nomenclature. *The Journal of the Royal Anthropological Institute*, 12, 23-44.

- Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *125*, 39-53.
- Blackmore, S. (1999). *The Meme Machine*. Oxford: Oxford University Press.
- Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.
- Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). “Bouba” and “kiki” in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners. *Cognition* *126*(2), 165-172.
- Brighton, H., & Kirby, S. (2006). Understanding linguistic evolution by visualising the emergence of topographic maps. *Artificial Life*, *12*, 229-242.
- Brighton, H., Kirby, S., & Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), *Language Origins: Perspectives on Evolution* (pp. 291-309). Oxford: Oxford University Press.
- Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, *2*, 177-226.
- Brown, R. W., Black, A. H., and Horowitz, A. E. (1955). Phonetic symbolism in natural languages. *Journal of Abnormal and Social Psychology*, *50*, 388–393.
- Cantrell, L. M., & Smith, L. B. (2013). Set size, individuation, and attention to shape. *Cognition*, *126*(2), 258-267.
- Campbell, L. (1998). *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Campbell, L., & Poser, W. J. (2008). *Language Classification*. Cambridge: Cambridge University Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, 15, 17–29.

Childs, G. T. (1994). African ideophones. In L. Hinton, J. Nichols, and J. J. Ohala (Eds.), *Sound Symbolism* (pp. 178-206). Cambridge: Cambridge University Press.

Chomsky, N. (1972). *Language and Mind* (2<sup>nd</sup> ed.). New York: Harcourt, Brace, & World.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.

Chomsky, N. (2005). Three factors in language design. *Linguistic Enquiry*, 36(1), 1-22.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 489–558.

Chu, M., and Kita, S. (2008). Spontaneous gestures during mental rotation tasks: insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology General*, 137, 706–723.

Claidière, N., Kirby, S., & Sperber, D. (2012). Effect of psychological bias separates cultural from biological evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(51), E3526.

Claidière, N., Smith, K., Kirby, S., & Fagot, J. (2014). Evolution of a systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society B*, 281, 20141541. <http://dx.doi.org/10.1098/rspb.2014.1541>

- Corballis, M. C. (2009). The evolution of language. *Annals of the New York Academy of Sciences*, 1156, 19-43.
- Cornish, H., Tamariz, M., & Kirby, S. (2010). Complex adaptive systems and the origins of adaptive structure: What experiments can tell us. *Language Learning*, 59: Supplement 1, 187-205.
- Croft, W. (1990). *Typology and Universals*. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.
- Croft, W. (2003). Typology. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (pp.434-440). London: Nature Publishing.
- Culbertson, J. & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, 6:1964. doi: 10.3389/fpsyg.2015.01964
- Cuskley, C. (2013). Mappings between linguistic sound and motion. *Public Journal of Semiotics*, 5(1), 39-62.
- Cuskley, C., & Kirby, S. (2013). Synaesthesia, cross-modality, and language evolution. In J. Simner & E. M. Hubbard (Eds.), *Oxford Handbook of Synaesthesia* (pp. 869-907). Oxford: Oxford University Press.
- Cuskley, C., Simner, J., & Kirby, S. (2015). Phonological and orthographic influences in the kiki-bouba effect. *Psychological Research*, doi: 10.1007/s00426-015-0709-2.
- Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London: John Murray.

- Davis, R. (1961). The fitness of names to drawings. A cross-cultural study in Tanganyika. *British Journal of Psychology*, 52, 259–268.
- DeLoache, J. S. (2002). Early development of the use of symbolic artifacts. In U. Goswami (Ed.), *Blackwell Handbook of Childhood Cognitive Development* (pp. 206–226). Malden, MA: Blackwell.
- DeLoache, J. S., & Burns, N. M. (1994). Early understanding of the representational function of pictures. *Cognition*, 52, 83-110.
- de Saussure, F. *Cours de linguistique générale*. Paris: Payot, 1916.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2014.00781
- Diffloth, G. (1972). The notes on expressive meaning. In P. M. Peranteau, J. N. Levi, and G. C. Phares (Eds.), *Papers from the Eighth Regional Meeting of Chicago Linguistic Society* (pp. 440-447) Chicago: Chicago Linguistic Society.
- Dingemanse, M. (2011). Ezra Pound among the Mawu: Ideophones and iconicity in Siwu. In P. Michelucci, O. Fischer, & C. Ljungberg (Eds.), *Semblance and Signification* (pp. 39-54). Amsterdam: John Benjamins.
- Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, 6, 654-672.
- D’Onofrio, A. (2014). Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm. *Language and Speech*, 57(3), 367-393.
- Drijvers, L., Zaadnoordijk, L., & Dingemanse, M. (2015). Sound-symbolism is disrupted in dyslexia: implications for the role of cross-modal abstraction processes. In D. Noelle,

- R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, & C. D. Jennings (Eds.) *Proceedings of the 37<sup>th</sup> Annual Meeting of the Cognitive Science Society* (pp.602-607). Austin, TX: Cognitive Science Society.
- Emmorey, K. (2014). Iconicity as structure mapping. *Philosophical Transactions of the Royal Society B*, 369, 20130301.
- E-Prime 2.0 software, Psychology Software Tools, Pittsburgh, PA.
- Esper, E. A. (1966). Social transmission of an artificial language. *Language*, 42(3), 575-580.
- Evans, N., & Levinson, S. (2009). The myth of language universals. *Behavioral and Brain Sciences*, 32(5) 429-448.
- Fantz, R. L., (1965). Visual perception from birth as shown by pattern selectivity. *Annals of the New York Academy of Sciences*, 118, 793–814.
- Fitch, W. T., & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings of the Royal Society B*, 268, 1669-1675.
- Folven R. J., Bonvillian, J. D. (1991). The transition from nonreferential to referential language in children acquiring American Sign Language. *Developmental Psychology*, 27, 806-816.
- Fodor, J. A. (1980). On the impossibility of acquiring “more powerful” structures. In M. Piattelli-Palmarine (Ed.), *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky* (pp. 142-149). Cambridge, MA: Harvard University Press.
- Fort, M., Martin, A., & Peperkamp, S. (2015). Consonants are more important than vowels in the kiki-bouba effect. *Language and Speech*, 58(2), 247-266.

- Frishberg, N. (1975). Arbitrariness and iconicity: historical change in American Sign Language. *Language*, 51, 676–710.
- Furman, R., Küntay, A. C., & Özyürek A. (2014). Early language-specificity of children's event encoding in speech and gesture: evidence from caused motion in Turkish. *Language, Cognition, and Neuroscience*, 29(5), 620-634.
- Gabry, J. and Goodrich, B. (2016). *rstanarm: Bayesian Applied Regression Modeling via Stan*. R package version 2.9.0-3. <http://CRAN.R-project.org/package=rstanarm>
- Galantucci, B., & Garrod, S. (2011). Experimental semiotics: A review. *Frontiers in Human Neuroscience*, 5, doi: 10.3389/fnhum.2011.00011.
- Galhardo, R. S., Hastings, P. J., & Rosenberg, S. M. (2007). Mutation as a Stress Response and the Regulation of Evolvability. *Critical Reviews in Biochemistry and Molecular Biology*, 42(5), 399–435.
- Garrod, S., Fay, N., Lee, J., Oberlander, J., & Macleod, T. (2007). Foundations of representation: where might graphical symbol systems come from? *Cognitive Science*, 31, 961–987.
- Gasser, N., Sethuraman, N., & Hockema, S. (2010). Iconicity in expressives: An empirical investigation. In S. Rice & J. Newman (Eds.), *Experimental and Empirical Methods in the Study of Conceptual Structure, Discourse, and Language* (pp. 163-180). Stanford, CA: CSLI Publications.
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), 1360-83.

- Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S.A. Kuczaj, II (Ed.), *Language development* (Vol. 2, pp. 301–334). Hillsdale, NJ: Erlbaum.
- Gentner, D. (2010). Bootstrapping the mind: analogical processes and symbol systems. *Cognitive Science*, 34, 752-775.
- Givón, T. (1985). Iconicity, isomorphism, and non-arbitrary coding in syntax. In J. Haiman (Ed.), *Iconicity in Syntax* (pp. 187-220). Amsterdam: John Benjamins.
- Givón, T. (1991). Isomorphism in the grammatical code: cognitive and biological considerations. *Studies in Language*, 15(1), 85-114.
- Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1, 1–55.
- Gluga, T., & Csibra, G. (2009). One-year-old infants appreciate the referential nature of deictic gestures and words. *Psychological Science*, 20(3), 347-353.
- The GNU Image Manipulation Program team, GIMP 2.8.0, [www.gimp.org](http://www.gimp.org), 1997-2012, retrieved on 02.07.2012.
- Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: the role of temporal synchrony between verbal labels and gestures. *Child Development*, 71(4), 878-894.
- Goldin-Meadow, S. (2014). Widening the lens: what the manual modality reveals about language, learning, and cognition. *Philosophical Transactions of the Royal Society B*, 369, 20130295.

- Goldin-Meadow, S., Goodrich, W., Sauer, E., & Iverson, J. (2007). Young children use their hands to tell their mothers what to say. *Developmental Science* 10(6), 778-785.
- Gontier, N. (2011). Depicting the tree of life: the philosophical and historical roots of evolutionary tree diagrams. *Evolution: Education and Outreach*, 4, 515-538.
- Goodrich, W., & Kam, C. L. H. (2009). Co-speech gesture as input in verb learning. *Developmental Science*, 12(1), 81-87.
- Goodwyn, S. W., Acredolo, L. P., & Brown, C. A. (2000). Impact of symbolic gesturing on early language development. *Journal of Nonverbal Behavior*, 24(2), 81-103.
- Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of Language* (pp.73-113). Cambridge, MA: MIT Press.
- Gullberg, M., Hendriks, H., & Hickmann, M. (2008). Learning to talk and gesture about motion in French. *First Language*, 28(2), 200-236.
- Haiman, J. (1980). The iconicity of grammar: isomorphism and motivation. *Language*, 56, 515-540.
- Haiman, J. (1985). *Natural Syntax. Iconicity and Erosion. Cambridge Studies in Linguistics, Vol. 44*. Cambridge: Cambridge University Press.
- Hall, T. A. (2007). Segmental Features. In P. de Lacy (Ed.) *The Cambridge Handbook of Phonology* (pp.311-334). Cambridge: Cambridge University Press.
- Hamano, S. (1998). *The sound-symbolic system of Japanese*. Stanford, CA & Tokyo: CSLI & Kuroshio.

- Hauser, M. D., Chomsky, N., & Fitch W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298, 1569-1579.
- Hayes, B. (2009). *Introductory Phonology*. Malden, MA: Wiley-Blackwell.
- Hirsh-Pasek, K., Golinkoff, R. M., & Hollich, G. (1999). Trends and transitions in language development: Looking for the missing piece. *Developmental Neuropsychology*, 16(2), 139–162.
- Hockett, C. (1960). The origin of speech. *Scientific American*, 203, 89-97.
- Hurford, J. R. (2002). Expression/induction models of language evolution: dimensions and issues. In E. J. Briscoe (Ed.), *Linguistic Evolution through Language Transmission* (pp. 324-352). Cambridge: Cambridge University Press.
- Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B*, 369, 20130298.
- Imai, M. Kita, S., Nagumo, M., and Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, 109, 54–65.
- Iverson, J. M., Capirci, O., Longobardi, E., & Caselli, M. C. (1999). Gesturing in mother-child interactions. *Cognitive Development*, 14, 57-75.
- Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, 16(5), 367-371.
- Iwasaki, N., Vinson, D. P., & Vigliocco, G. (2007a). How does it hurt, kiri-kiri or siku-siku? Japanese mimetic words of pain perceived by Japanese speakers and English

- speakers. In M. Minami (Ed.) *Applying Theory and Research to Learning Japanese as a Foreign Language* (pp. 2-19). Newcastle: Cambridge Scholars Publishing.
- Iwasaki, N., Vinson, D. P., & Vigliocco, G. (2007b). What do English speakers know about gera-gera and yota-yota? A cross-linguistic investigation of mimetic words for laughing and walking. *Japanese Language Education around the Globe*, 17, 53–78.
- Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.
- Jeffreys, H. (1961). *Theory of probability (3rd ed.)*. Oxford, UK: Oxford University Press.
- Johnson, K. E., Younger, B. A., & Furrer, S.D. (2005). Infants' symbolic comprehension of actions modelled with two replicas. *Developmental Science*, 8(4), 299-318.
- Kantartzis, Imai, & Kita (2011). Japanese sound-symbolism facilitates word learning in English-speaking children. *Cognitive Science*, 35, 575-586.
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773-795.
- Kaye, J. (1989). *Phonology: A Cognitive View*. Hillsdale NJ: Lawrence Erlbaum Associates.
- Kelly, S. D., Özyürek, A., and Maris, E. (2010). Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21, 260–267.
- Kempe, V., Gauvrit, N., Forsyth, D. (2015). Structure emerges faster during cultural transmission in children than in adults. *Cognition*, 136, 247-254.
- Kendon, A. (2004). *Gesture: visible actions as utterance*. Cambridge UK: Cambridge University Press.

- Kewley-Port, D., Burkle, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to intelligibility for young normal-hearing and elderly hearing-impaired listeners. *Journal of the Acoustical Society of America*, 122(4), 2365-2375.
- Kirby, S. (1999). *Function, Selection, and Innateness: The Emergence of Language Universals*. Oxford: Oxford University Press.
- Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, M. Studdert-Kennedy, & J. R. Hurford (Eds.), *The Evolutionary Emergence of Language: Social Functions and the Origins of Linguistic Form* (pp. 303-323). Cambridge: Cambridge University Press.
- Kirby, S. (2002). Learning, bottlenecks, and the evolution of recursive syntax. In Briscoe, T. (Ed.), *Linguistic Evolution Through Language Acquisition: Formal and Computational Models* (pp.173-204). Cambridge: Cambridge University Press.
- Kirby, S., Cornish, H., Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31), 10681-10686.
- Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 104(12), 5241-5245.
- Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28, 108-114.

- Kirby, S., & Hurford, J. R. (1997). Learning, culture, and evolution in the origin of linguistic constraints. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life* (pp. 493-502). Cambridge, MA: MIT Press.
- Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the Evolution of Language* (pp. 121-148). London: Springer.
- Kirby, S., Smith, K., and Brighton, H. (2004). From UG to universals: linguistic adaptation through iterated learning. *Studies in Language*, 28(3), 587–607.
- Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. *Linguistics*, 35, 379-415.
- Kita, S. (2008). World-view of protolanguage speakers as inferred from semantics of sound symbolic words: a case of Japanese mimetics. In N. Masataka (Ed.), *The origins of language* (pp. 25 – 38). Tokyo: Springer.
- Köhler, W. (1929). *Gestalt Psychology*. New York: Liveright.
- Köhler, W. (1947). *Gestalt Psychology*, 2<sup>nd</sup> Ed. New York: Liveright.
- Koriat, A., & Levy, I (1977). The symbolic implications of vowels and of their orthographic representations in two natural languages. *Journal of Psycholinguistic Research*, 6(2), 93–104.
- Kovic, V., Plunkett K., and Westermann, G. (2010). The shape of words in the brain. *Cognition*, 114, 19–28.
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science* 6(3), 299-312.

- Ladefoged, P. (2001). *A Course in Phonetics* (4<sup>th</sup> ed.). Boston MA: Heinle and Heinle.
- Laing, C. E. (2014). A phonological analysis of onomatopoeia in early word production. *First Language*, 34(5), 387-405.
- Larson, E. J. (2004), *Evolution: The Remarkable History of a Scientific Theory*. New York: Modern Library.
- Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1–75.
- Levinson, S. C. (2000). *Presumptive Meanings*. Cambridge, MA: MIT Press.
- Liljencrants, J. & Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language*, 48(4), 839-862.
- Lizkowski, U. (2010). Deictic and other gestures in infancy: Deicticos y otros gestos en la infancia. *Accion Psicologica*, 7, 21–33.
- Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: a review of behavioural, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, 6(1246). doi:10.3389/fpsyg.2015.01246
- 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.
- Lockwood, G., Hagoort, P., & Dingemanse, M. (2016). How iconicity helps people learn new words: Neural correlates and individual differences in sound-symbolic bootstrapping. *Collabra*, 2(1), 1–15.

Lockwood, G., Tuomainen, J. (2015). Ideophones in Japanese modulate the P2 and late positive complex responses. *Language Sciences*, 6, 933.  
DOI:10.3389/fpsyg.2015.00933

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, 50, 315-353.

Maeda, T., and Maeda, K. (1983). *Yoji no goihattatsu no kenkyu [Investigation of a Child's Lexical Development]*. Tokyo: Musashino Shoin.

Maguire, M. J., Hirsh-Pasek, K., Golinkoff, R. M., Imai, M., Haryu, E., Vanegas, S., & Sanchez-Davis, B. (2010). A developmental shift from similar to language-specific strategies in verb acquisition: a comparison of English, Spanish, and Japanese. *Cognition*, 114, 299–319.

Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 384-394.

MATLAB 2012a, The MathWorks, Inc., Natick, Massachusetts, United States

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Science*, 9(3), 316-322.

McGregor, W. (2001). Ideophones as the source of verb in Northern Australian languages. In F. K. E. Voeltz and C. Kilian-Hatz (Eds.), *Ideophones* (pp. 205-211). Amsterdam: John Benjamins.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Language and Thought*. Chicago: University of Chicago Press.

- Meteyard, L., Stoppard, E., Snudden, D., Cappa, S. F. and Vigliocco, G. (2015) When semantics aids phonology: a processing advantage for iconic word forms in aphasia. *Neuropsychologia*, 76, 264-275.
- Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Brockman, W., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014), 176-182.
- Mikone, E. (2001). Ideophones in the Balto-Finnic Languages. In F. K. E. Voeltz and C. Kilian-Hatz (Eds.), *Ideophones* (pp. 223-233). Amsterdam: John Benjamins.
- Mitterer, H., Schuerman, W., Reinisch, E., Tufvesson, S. & Dingemanse, M. (2012). The limited power of sound symbolism. In *Proceedings of the 16<sup>th</sup> Annual Conference on Architectures and Mechanisms for Language Processing* (pp. 27). Rivades Garda, Italy.
- Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140, 325–347.
- Monaghan, P., Mattock, K., Davies, R., & Smith, A.C. (2015). Gavagai is as gavagai does: Learning nouns and verbs from cross-situational statistics. *Cognitive Science*, 39, 1099-1112.
- Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in word learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 38(5), 1152-1164.

- Monaghan, P., Shillcock, R. C., Christiansen, M., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B*, 369(1651). DOI: 10.1098/rstb.2013.0299
- Moore, C., Angelopoulos, M., & Bennett, P. (1999). Word learning in the context of referential and salience cues. *Developmental Psychology*, 35(1), 60–68.
- Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*. doi:10.1016/j.jmp.2015.11.001
- Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-2. <http://CRAN.R-project.org/package=BayesFactor>
- Mumford, K. H., & Kita, S. (2014). Children use gesture to interpret novel verb meanings. *Child Development*, 85(3), 1181-1189.
- Namy, L. L. (2008). Recognition of iconicity doesn't come for free. *Developmental Science*, 11(6), 841-846.
- Newmeyer, F. J. (1992). Iconicity and generative grammar. *Language*, 68, 756–796.
- Nicoladis, E. (2002). Some gestures develop in conjunction with spoken language development and others don't: evidence from bilingual preschoolers. *Journal of Nonverbal Behavior*, 26(4), 241-266.
- Nielsen, A. K. S., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology*, 62(2), 115-124.

- Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, 4(2), 115-125.
- Nielsen, A. K. S., & Rendall, D. (2013). Parsing the role of consonants versus vowels in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology*, 67(2), 153-163.
- Nuckolls, J. (1996). *Sounds Like Life*. New York: Oxford University Press.
- Nygaard, L. C., Herold, D. S., and Namy, L. L. (2009). The semantics of prosody: listeners' use of acoustic correlates to word meaning. *Cognitive Science*, 33, 127–146.
- Ocelli, V., Esposito, G., Venuti, P., Arduino, G. M., & Zampini, M. (2013). The takete-maluma phenomenon in autism spectrum disorders. *Perception*, 42, 233-241.
- Oda, H. (2000). *An Embodied Semantic Mechanism for Mimetic Words in Japanese*. Bloomington, IN: Indiana University.
- Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, 41, 1 - 16.
- Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.) *Sound Symbolism* (pp. 325-347). Cambridge: Cambridge University Press.
- Orlansky, M. D., & Bonvillian, J. D. (1984) The role of iconicity in early sign language acquisition. *Journal of Speech and Hearing Disorders*, 49, 287– 292.

- Owren, M. J., & Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *Journal of the Acoustical Society of America*, 119(3), 1727–1739.
- Özçalışkan, Ş., Gentner, D., & S. Goldin-Meadow (2014). Do iconic gestures pave the way for children's early verbs? *Applied Linguistics*, 35, 1143-1162.
- Özçalışkan, Ş., & Goldin-Meadow, S. (2009). When gesture-speech combinations do and do not index language change. *Language and Cognitive Processes*, 24(2), 190-217.
- Özçalışkan, Ş., and Goldin-Meadow, S. (2011). Is there an iconic gesture spurt at 26 months? In G. Stam & M. Ishino (Eds.), *Integrating Gestures: The Interdisciplinary Nature of Gestures* (pp. 163-174). Amsterdam: John Benjamins.
- Ozturk, O., Krehm, Madelaine, & Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound-shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, 114, 173-186.
- Özyürek A., Kita, S., Allen, S., Brown A., Furman, R., & Ishizuka, T. (2008). Development of cross-linguistic variation in speech and gesture: motion events in English and Turkish. *Developmental Psychology*, 44(4), 1040-1054.
- Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound-symbolism: a study using the implicit association test. *Experimental Brain Research*, 220(3), 319-333.
- Peirce, C.S. (1931–35, 1958). *Collected Papers of Charles Sanders Peirce*, vols. 1–6, (1931–35) C. Hartshorne and P. Weiss (Eds.), vols. 7–8 (1958) A. W. Burks (Ed.). Cambridge, MA: Harvard University Press.

- Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society Open Science* 2, 150152.  
<http://dx.doi.org/10.1098/rsos.150152>
- Perniss, P., Lu, J., Morgan, G., & Vigliocco, G. (under revision). Mapping language to the world: The role of iconicity in the sign language input. *Submitted to Developmental Science*.
- Perniss, P., Thompson, R. L., & G. Vigliocco (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Language Science*, 1(227). DOI: 10.3389/fpsyg.2010.00227
- Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: From a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B*. 369(1651), 20130300, 1-13.
- Perry, L. K., Perlman, M., Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLoS* 1, 10(9): e0137147.  
DOI:10.1371/journal.pone.0137147
- Pitcher, B. J., Mesoudi, A., & McElligott, A. G. (2013). Sex-biased sound symbolism in English-language first names. *PLoS ONE*, 8(6), e64825.  
doi:10.1371/journal.pone.0064825
- Pinker, S. & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13, 707–727.
- Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. New York: William Morrow and Company.

Plato (360 B.C.). *Cratylus* (B. Jowett, Trans.). The Internet classics archive. URL: <http://classics.mit.edu/index.html>, retrieved 05/16/2016.

Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., & Hennon, E. A. (2006). The birth of words: Ten month-olds learn words through perceptual salience. *Child Development, 77*(2), 266–280.

Qualtrics (2015). Qualtrics Inc., Provo, Utah, USA. URL <http://www.qualtrics.com>

Quine, W. V. (1960). *Word and Object*. Cambridge, MA: MIT Press.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.

Ramachandran, V., & Hubbard, E. (2001). Synaesthesia: A window into perception, thought, and language. *Journal of Consciousness Studies, 8*(1), 3-34.

Rastle, K. & Brysbaert, M. (2006). Masked phonological priming effects in English: Are they real? Do they matter? *Cognitive Psychology, 53*, 97-145.

Revill, K. P., Namy, L. L., De Fife, L. C., & Nygaard, L. C. (2014). Cross-linguistic sound symbolism and cross modal correspondence: evidence from fMRI and DTI. *Brain and Language, 128*, 18–24.

Richerson, P. J., & Boyd R. (2004). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.

Ricker, W. E. (1981). Changes in the Average Size and Average Age of Pacific Salmon. *Canadian Journal of Fisheries and Aquatic Sciences, 38*, 1636-1656.

- Roach, P. (2004). British English: Received Pronunciation. *Journal of the International Phonetic Association*, 34(2), 239–245.
- Rogers, D. S. & Ehrlich, P. R. (2008). Natural selection and cultural rates of change. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9), 3416-3420.
- Ross, J.R. (1967). *Constraints on Variables in Syntax*. Doctoral dissertation, MIT.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). *Psychonomic Bulletin & Review*, 16(2), 225-237.
- Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, 323, 951-953.
- Rowe, M., L., Özçalışkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: gesture's role in predicting language development. *First Language*, 28(2), 182-199.
- Saji, N. & Imai, M. (2013). Onomatopoe kenkyu no shatei – chikadzuku oto to imi [Sound Symbolism and Mimetics]. In K. Shinohara & R. Uno (Eds.), *Goishutoku ni okeru ruizousei no kouka no kentou* (pp. 151-166). Tokyo: Hituji Syobo.
- Samuelson, L. K., & Smith, L. B. (2005). They Call It Like They See It: Spontaneous Naming and Attention to Shape. *Developmental Science*, 8(2), 182-198.
- Sandler, W. (2009). Symbiotic symbolization by hand and mouth in sign language. *Semiotica*, 174, 241–275.
- Schourup, L. (1993). Nihongo no kakikotoba-hanashikotoba ni okeru onomatopoe no bunpu nitsuite [on distribution of onomatopoeias in spoken and written Japanese]. In H.

- Takehi and I. Tamori (Eds.), *Onomatopia: Giongo-gitaigo no Rakuen [Onomatopoeia: Paradise of Mimetics]* (pp. 48-55). Tokyo: Keiso Shobo.
- Schultze-Berndt, E. (2001). Ideophone-like characteristics of uninflected predicates in Jaminjung (Australia). In F. K. E. Voeltz and C. Kilian-Hatz (Eds.), *Ideophones* (pp. 355-373). Amsterdam: John Benjamins.
- Scott-Phillips, T., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Science*, 14(9), 411-417.
- Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113, 226-233.
- Searle, J. R. (2002). Indeterminacy, empiricism, and the first person. In *Consciousness and Language*. Cambridge: Cambridge University Press.
- Shintel, H., Nusbaum, H. C., and Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language*, 55, 165–177.
- Siskind, J.M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61, 39–61.
- Smith, K. & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116, 444-449.
- Smith, L.B. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 51–80). Oxford, England: Oxford University Press.
- Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558-1568.

- Snyder, W. (2011). Children's Grammatical Conservatism: Implications for syntactic theory [Plenary Address]. In N. Danis, K. Mesh & H. Sung (Eds.), *BUCLD 35: Proceedings of the 35th annual Boston University Conference on Language Development, Volume I*, 1-20. Somerville, MA: Cascadilla Press.
- Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, and Psychophysics*, 73, 971-995.
- Stan Development Team (2015). *Stan: A C++ Library for Probability and Sampling, Version 2.8.0*. URL <http://mc-stan.org/>.
- Stan Development Team (2015). *Stan Modeling Language User's Guide and Reference Manual, Version 2.6.1*. URL <http://mc-stan.org/>.
- Stanfield, C., Williamson, R., & Özçalışkan, Ş. (2014). How do children understand gesture-speech combinations with iconic gestures? *Journal of Child Language*, 41, 462-471.
- Sterelny, K. (2012). Language, gesture, skill: The co-evolutionary foundations of language. *Philosophical Transactions of the Royal Society of London Series B*, 367(1599), 2141-2151.
- Striano, T., Rochat, P., & Legerstee, M. (2003). The role of modelling and request type on symbolic comprehension of objects and gestures in young children. *Journal of Child Language*, 30, 27-45.
- Tanz, C. (1971) Sound Symbolism in Words Relating to Proximity and Distance. *Language and Speech*, 14, 266– 276.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279-1285.

Testable: The web platform for creating, running and sharing behavioural experiments (2016). [www.testable.org](http://www.testable.org)

Theisen, C.A., Oberlander, J., & Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. *Interaction Studies*, *11*, 14–32.

Theisen-White, C., Kirby, S., & Oberlander, J. [2011] Integrating the horizontal and vertical cultural transmission of novel communication systems. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Boston*. 956-961.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(16), 4530-4535.

Thompson, R. L., Vinson, D. P., and Vigliocco, G. (2009). The link between form and meaning in American sign language: lexical processing effects. *Journal of Experimental Psychology: Language, Memory, and Cognition*, *35*, 550–557.

Thompson, R. L., Vinson, D. P., and Vigliocco, G. (2010). The link between form and meaning in British sign language: lexical processing effects in a phonological decision task. *Journal of Experimental Psychology: Language, Memory, and Cognition*, *36*, 1017–1027.

Thompson, R.L., Vinson, D.P., Woll, B., & Vigliocco, G. (2012). The road to language learning is iconic. *Psychological Science*, *23*(12), 1443-1448.

- Tolar, T. D., Lederberg, A. R., Gokhale, S., Tomasello, M. (2008). The development of the ability to recognise the meaning of iconic signs. *Journal of Deaf Studies and Deaf Education*, 13(2), 225-240.
- Tomasello, M. (2000). Perceiving intentions and learning words in the second year of life. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 111–128). Cambridge, England: Cambridge University Press.
- Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.
- Ullian, R. 1978. Size-sound symbolism. In J. H. Greenberg, C. A. Ferguson, and E. A. Moravcsik (Eds.), *Universals of human language, Vol. 2: Phonology* (pp.527-568). Stanford, CA: Stanford University Press.
- Urban, M. (2011). Conventional sound symbolism in terms for organs for speech”A cross-linguistic study. *Folia Linguistica*, 45(1), 199-213.
- Verhoef, T., Kirby, S., & De Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, 43, 57-68.
- Verhoef, T., Kirby, S., de Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive Science*, 40, 1969-94.
- Vigliocco, G., and Kita, S. (2006). Language specific effects of meaning, sound and syntax: implications for models of lexical retrieval in production. *Language and Cognitive Processes*, 21, 790–816.

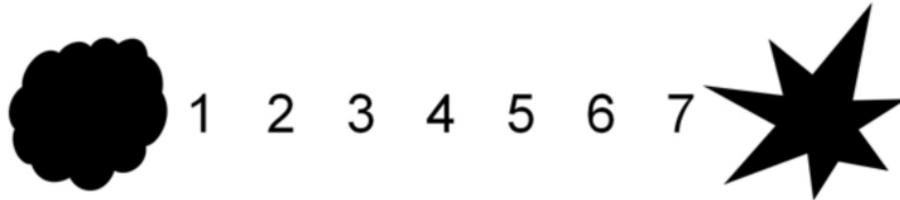
- Vinson, D., Thompson, R. L., Skinner, R., & Vigliocco, G. (2015). A faster path between meaning and form? Iconicity facilitates sign recognition and production in British Sign Language. *Journal of Memory and Language*, *82*, 56-85.
- Vogt, P., & Smith, A. D. M. (2005). Learning colour words is slow: A cross-situational learning account. *Behavioral and Brain Sciences*, *28*, 509–510.
- Wagenmakers, E. J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin & Review*, *14*(5), 779-804.
- Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, *21*(1), 21-25.
- Watson, R. L. (2001). A comparison of some Southeast Asian ideophones with some African ideophones. In F. K. E. Voeltz and C. Kilian-Hatz (Eds.), *Ideophones* (pp. 385-405). Amsterdam: John Benjamins.
- Waxman, S., Fu, X., Arunachalam, S., Leddone, E., Geraghty, K., & Song, H. (2013). Are nouns learned before verbs? Infants provide insight into a long-standing debate. *Child Development Perspectives*, *7*(3), 155-159.
- Westbury, C. (2005). Implicit sound symbolism in lexical access: evidence from an interference task. *Brain and Language*, *93*, 10–19.
- Wittgenstein, L. (1953/2009). *Philosophical Investigations*. Chichester: Blackwell.
- Xu, Y., Kelly, A., Smillie, C. (2013). Emotional expressions as communicative signals. In S. Hancil and D. Hirst (Eds.) *Prosody and Iconicity* (pp. 33-60). Amsterdam: John Benjamins.

- Yoshida, H. (2012). A cross-linguistic study of sound-symbolism in children's verb learning. *Journal of Cognition and Development, 13*(2), 232-265.
- Yu, C. and Ballard, D. H. (2007) A Unified Model of Early Word Learning: Integrating Statistical and Social Cues. *Neurocomputing, 70*(13-15), 2149-2165.
- Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(15), 414–420.
- Yu, C., & Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science, 14*(2), 165-180.
- Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: the baby's view is better. *Developmental Science, 16*(6), 959-966.



# Appendices

## Appendix 2.1: Ratings Scale and Instructions



Thank you for signing up to the study, and welcome!

In the following trials you will hear a series of speech recordings. Your job is to rate them.

You will **rate the sound on how round or spiky it seems** to you. This may seem strange, but if you think about it some sounds could be said to sound round (e.g. *wubo*) and others could be said to sound spiky (e.g. *ziki*). You will rate on a scale of 1-7, where **1 is the roundest, 7 is the spikiest, and 4 is in the middle, like on the scale above.**

There are no right or wrong answers here - just go with your instinct.

**You will rate by pressing a keyboard key, 1-7. There is no need to press ENTER; the next trial will start automatically.**

You will rate 200-300 recordings in this experiment. Each trial should be very quick, and the study should take roughly 30 minutes.

**IMPORTANT - This study is based on sound! Please ensure that you are in a quiet place and have functioning headphones/earphones/speakers. EACH SOUND WILL ONLY PLAY ONCE, AT THE BEGINNING OF THE TRIAL - MAKE SURE YOU'RE LISTENING!!!**

*An Example of the instructions issued to participants before the ratings study*

## Appendix 2.2: Complete Table of Syllable Scores

	p	b	m	f	v	θ	ð	t	d	s	z	n	l	tʃ	dʒ	ʃ	ʒ	ʤ	j	k	g	ŋ	w	h	Mean
i	5.30	4.16	4.28	4.60	5.00	4.84	4.94	5.68	4.76	4.33	4.84	4.00	3.94	4.96	4.26	4.64	4.74	3.60	3.75	5.84	5.04	3.36	3.88	3.37	4.51
ɪ	4.52	3.59	3.92	4.30	4.22	4.00	4.70	5.60	4.67	4.54	4.81	4.04	4.37	4.85	4.44	4.32	5.22	4.04	3.67	5.63	4.93	4.28	3.48	4.70	4.45
e	4.30	3.92	3.88	4.33	4.59	4.32	4.20	4.92	4.26	4.17	4.84	3.92	3.92	5.04	4.59	4.22	4.32	3.33	3.24	5.19	4.72	4.00	3.44	4.22	4.25
ɛ	4.72	4.33	4.11	4.73	4.96	4.30	3.44	5.07	4.28	4.28	5.33	4.12	4.15	5.32	4.59	4.44	4.44	3.89	4.12	4.81	5.04	3.88	3.52	3.92	4.41
æ	4.44	3.96	3.78	4.24	4.72	4.20	4.56	5.00	4.33	3.74	5.16	3.38	3.48	4.78	4.48	4.67	4.28	3.97	3.65	5.76	4.42	4.20	3.42	4.41	4.29
ɜ	4.88	3.81	2.93	4.28	3.89	4.64	3.81	4.81	4.16	4.15	4.64	3.74	3.81	5.36	4.30	3.72	4.24	3.32	3.24	4.80	4.70	3.48	3.20	3.89	4.08
ə	4.74	4.56	3.85	3.70	4.56	3.81	4.63	5.36	4.15	4.56	5.15	4.04	4.00	5.24	5.08	4.60	4.44	3.67	4.00	5.48	4.36	3.59	3.60	3.92	4.38
ʌ	4.88	4.28	3.76	4.32	4.48	4.78	3.92	5.60	4.58	4.26	4.67	4.22	3.37	4.74	4.89	4.04	4.22		3.88	5.60	5.00	3.96	3.37	4.63	4.41
ɑ	4.16	4.00	3.36	4.11	4.48	3.64	4.28	5.08	4.20	4.16	4.84	4.04	3.74	5.24	4.36	4.12	5.19	3.74	4.28	6.24	4.16	4.30	3.07	4.08	4.29
ɑ:	4.12	3.26	3.12	3.93	4.24	3.16	4.08	4.40	3.92	3.41	4.20	3.52	2.72	4.20	4.00	3.88	4.19	3.48	2.96	5.08	4.16	2.92	2.36	2.63	3.66
ɒ	3.56	3.84	3.74	3.89	4.36	4.32	3.78	4.68	3.93	4.26	4.74	3.81	3.33	5.04	4.64	4.41	4.19		3.59	5.04	4.52	3.92	3.00	3.48	4.09
ɔ	3.36	3.78	3.30	4.04	4.30	4.08	3.88	5.16	4.63	3.72	4.24	3.59	3.60	4.74	4.56	3.88	4.48	3.74	3.04	4.70	4.72	3.33	2.85	3.76	3.98
ɔ:	3.28	2.89	2.93	3.32	3.33	3.56	4.12	4.00	3.48	3.36	4.04	2.78	2.50	4.36	3.56	3.06	2.60	2.52	2.76	4.34		2.78	1.92	2.72	3.22
u	3.44	3.37	2.20	3.52	3.41	4.07		4.76	3.31	3.30	4.63	3.11	3.30	4.37	4.37	3.93	3.88	2.92	3.33	4.72	3.63	2.74	2.64	2.00	3.52
ʊ	4.84	3.96	3.52	4.12	4.17	4.81	3.81	5.36	4.36	4.42	4.00	3.92	3.67	4.81	4.92	4.19	4.33		3.80	5.60	4.48	3.92	2.41	4.33	4.25
eɪ	4.5	4.2	3.5	4.1	3.8	3.9	4.0	4.1	4.0	4.1	4.0	3.6	3.5	4.1	4.6	4.1	4.0	3.4	3.2	5.4	3.6	4.2	3.0	4.0	4.0

	6	8	2	5	5	6	6	5	0	5	4	7	9	5	4	2	4	3	4	8	7	8	0	4	0
ai	4.8 4	4.0 8	3.5 6	4.2 6	4.5 9	3.9 3	4.2 6	4.6 3	4.3 6	4.2 6	4.6 4	3.9 2	3.7 6	4.7 1	4.8 0	4.0 0	3.9 3	3.8 5	3.4 8	5.4 8	3.9 6	3.7 2	3.7 4	3.8 4	4.1 9
ɔi	3.8 5	3.4 1	3.3 6	3.4 1	4.0 4	3.8 1	3.5 6	4.1 9	3.5 2	3.5 9	4.3 7	3.7 4	3.2 0	4.5 6	4.0 8	3.7 4	3.1 3	3.7 0	2.9 3	4.4 1	4.1 6	3.3 6	2.7 8	3.1 1	3.6 7
ɪə	4.1 1	3.8 0	3.4 1	4.0 7	4.1 9	3.7 0	4.6 0	4.3 3	4.0 8	4.1 6	5.0 0		3.6 9	5.2 0	4.9 2	3.7 6	4.2 2	3.8 0	3.4 2	5.2 8	4.4 8	3.8 8	3.5 9	4.1 1	4.1 7
eə	3.8 1	3.5 6	3.7 4	3.5 6	4.4 8	3.9 2		4.1 9	4.5 6	3.8 8	4.8 8	3.7 8	3.4 1	4.7 0	4.1 3	3.9 3	4.0 0	3.5 6	3.1 5	5.0 7	4.3 0	3.5 6	3.2 7	3.8 1	3.9 7
ʊə	3.6 0	3.6 3	2.7 6	4.0 4	4.0 0	3.7 8		3.8 1	3.5 9	4.0 7	4.5 9	3.1 6		4.4 0	4.0 4	3.2 0		3.7 4	2.8 1	4.6 7	4.3 6	2.7 2	2.8 4	3.4 8	3.6 8
ɪʊ	3.2 0	2.8 1	2.7 0	2.6 4	3.3 7	3.4 8	3.7 0	3.6 7	3.6 4	2.8 4	4.4 4	3.1 5	2.9 2	4.2 6	3.4 8	3.2 4	3.4 8	2.9 6	2.9 2	3.9 6	3.8 9	3.2 2	2.4 8	2.4 4	3.2 9
aʊ	3.8 4	3.4 8	3.6 8	3.5 2	3.6 4	3.7 2	3.4 0	4.8 9	3.7 4	4.4 4	4.7 7	3.4 8	3.5 9	4.5 2	4.4 8	4.3 3	4.7 4	3.4 3	3.7 0	5.0 4	4.8 3	3.6 3	3.0 4	3.7 8	3.9 9
əʊ	4.0 4	4.0 8	2.8 8	3.9 6	3.6 4	3.7 4	3.8 1	4.1 5	3.6 4	3.7 4	4.6 8	3.4 8	2.6 0	3.8 0	4.3 6	3.6 8	3.4 8	3.2 8	2.7 6	4.3 7	4.1 6	3.2 8	2.8 9	3.1 6	3.6 5
Me an	4.1 8	3.7 9	3.4 3	3.9 6	4.1 9	4.0 2	4.0 7	4.7 3	4.0 9	3.9 9	4.6 5	3.6 8	3.5 1	4.7 2	4.4 2	4.0 0	4.1 6	3.5 2	3.4 1	5.1 1	4.4 2	3.6 0	3.0 7	3.6 6	4.0 2

*The complete set of syllable scores for the CV syllables normed in the study. Each column is headed by the consonant common to all syllables appearing in that column, and each row by the vowel common to all syllables in that row. The first two digits of the number in each cell appear in the first line, the third appears on the second line.*

*Consonants are arranged by place of articulation, from lips to glottis, and within that by ascending sonority. Monophthongs are arranged – roughly speaking – moving through the vowel space in a u shape from [i] to [u], with lax vowels appearing after their tensed equivalents. Diphthongs are at the bottom of the table arranged by the same scheme, applying first to target position and within that to start position. Means of means appear in the bottom row (by consonant) and the rightmost column (by vowel). The bottom rightmost cell is the mean of all means. Blacked-out cells represent syllables inadvertently omitted from the study.*

*Values of above four represent roundness, and values below four represent spikiness. Cells are colour coded according to the distance from the overall mean, with spiky values being redder, and round values being greener. This provides a useful visual shortcut for gauging the overall roundness/spikiness of a particular phoneme: look down its column/across its row to see how vivid and consistent its colour scheme is.*

## Appendix 2.3: Standard Deviations of Ratings

	p	b	m	f	v	θ	ð	t	d	s	z	n	l	tʃ	dʒ	ʃ	ʒ	ʒ	j	k	g	ŋ	w	h	Me an
i	1.7 7	1.6 5	1.6 7	1.2 9	1.4 7	1.3 7	1.3 3	1.1 8	1.5 9	1.7 5	1.2 1	1.4 1	1.3 9	1.4 5	1.5 6	1.3 2	1.5 6	1.4 4	1.7 0	1.3 7	1.7 5	1.7 5	1.6 9	1.3 6	1.5 0
ɪ	1.8 3	1.6 5	1.6 3	1.3 5	1.6 7	1.4 9	1.4 9	1.3 2	1.5 7	1.3 0	1.4 9	1.4 3	1.5 7	1.6 3	1.3 7	1.2 8	1.3 4	1.5 4	1.7 8	1.4 2	1.4 4	1.7 7	1.5 0	1.4 4	1.5 1
e	1.2 0	1.1 9	1.4 2	1.3 3	1.4 2	1.2 2	1.3 2	1.4 1	1.3 5	1.4 3	1.4 9	1.4 7	1.5 5	1.3 4	1.4 2	1.8 7	1.0 7	1.4 1	1.3 3	1.2 1	1.2 8	1.7 5	1.5 8	1.4 8	1.4 0
ɛ	1.3 7	1.4 9	1.5 0	1.4 0	1.1 4	1.5 4	1.1 2	1.4 1	1.4 9	1.2 4	1.3 6	1.6 4	1.4 3	1.3 1	1.5 3	1.3 3	1.4 2	1.4 2	1.5 1	1.8 0	1.2 1	1.3 3	1.5 0	1.5 5	1.4 2
æ	1.5 8	1.5 3	1.7 8	1.3 3	1.2 4	1.4 1	1.3 3	1.3 9	1.7 1	1.7 0	1.3 1	1.5 5	1.6 1	1.4 5	1.5 3	1.4 4	1.0 6	1.4 8	1.6 5	1.2 3	1.5 0	1.8 5	1.9 6	1.8 9	1.5 2
ɜ	1.6 2	1.7 1	1.3 0	1.3 7	1.6 3	1.3 2	1.4 2	1.6 2	1.1 8	1.3 5	1.5 0	1.3 8	1.7 1	1.2 5	1.6 6	1.0 2	1.3 3	1.4 9	1.4 8	1.6 1	1.5 9	1.5 3	1.5 8	1.4 8	1.4 6
ə	1.8 7	1.5 6	1.7 5	1.5 4	1.4 2	1.3 3	1.2 8	1.4 4	1.6 8	1.3 7	1.5 4	1.3 7	1.6 6	1.5 1	1.2 2	1.3 5	1.4 2	1.5 7	1.8 4	1.6 4	1.6 0	1.6 9	1.7 8	1.5 5	1.5 4
ʌ	1.6 9	1.7 4	1.7 6	1.4 1	1.5 8	1.4 2	1.3 2	1.0 0	1.5 5	1.5 6	1.2 4	1.3 7	1.7 8	1.5 3	1.4 5	1.6 5	1.5 8	1.3 8	1.5 9	1.2 9	1.2 4	1.7 0	1.5 0	1.9 0	1.5 2
ɑ	1.7 5	1.6 3	1.4 4	1.4 0	1.7 2	1.3 8	1.3 1	1.6 1	1.2 6	0.9 4	1.5 2	1.4 8	1.6 1	1.3 3	1.3 5	1.4 5	1.2 1	1.3 8	1.8 6	0.9 7	1.4 9	1.6 4	1.7 5	1.3 2	1.4 5
ɑ:	1.6 2	1.5 6	1.1 7	1.4 7	1.5 1	1.3 1	1.0 8	1.9 8	1.7 3	1.3 7	1.6 1	1.5 0	1.1 7	1.2 9	1.2 6	1.5 4	1.6 2	1.3 4	1.5 9	1.5 0	1.4 3	1.3 5	1.3 5	1.1 7	1.4 4
ɒ	1.6 5	1.6 0	1.5 1	1.4 2	1.5 5	1.1 8	1.6 3	1.3 1	1.7 3	1.6 5	1.7 5	1.3 0	1.5 2	1.4 6	1.4 4	1.5 0	1.4 4	1.3 4	1.5 5	1.6 8	1.2 9	1.5 8	1.4 4	1.6 6	1.5 1
ɔ	1.6	1.5	1.5	1.3	1.5	1.5	1.2	1.2	1.3	1.6	1.6	1.5	1.7	1.6	1.4	1.4	1.5	1.3	1.4	1.6	1.7	1.4	1.8	1.3	1.5

	0	3	1	1	9	0	4	8	9	5	9	5	3	3	2	2	8	8	8	8	4	1	3	3	2
ɔ:	1.7 7	1.4 2	1.3 8	1.4 4	1.5 4	1.1 2	1.0 1	1.7 1	1.4 5	1.4 7	1.6 0	1.2 5	1.0 2	1.7 8	1.6 7	1.4 1	1.1 2	1.2 3	1.4 2	1.7 8		1.5 8	1.1 5	1.1 7	1.4 1
u	1.7 8	1.7 6	1.0 8	1.4 8	1.6 7	1.5 9		1.3 3	1.5 9	1.4 4	1.8 8	1.4 0	1.3 8	1.6 4	1.5 2	1.5 9	1.5 4	1.2 9	1.5 4	1.7 2	1.5 5	1.3 2	1.6 0	1.2 2	1.5 2
ʊ	1.5 2	1.8 5	1.4 5	1.7 6	1.5 5	1.5 7	1.3 3	1.4 1	1.6 6	1.5 0	2.0 4	1.9 6	1.5 9	1.6 3	1.3 5	1.3 3	1.5 7		1.6 6	1.6 3	1.6 7	1.7 1	1.6 7	1.7 1	1.6 1
eɪ	1.4 7	1.2 1	1.1 9	1.1 0	1.1 7	1.0 6	1.3 0	1.5 1	1.6 6	1.4 3	1.1 7	1.5 2	1.6 2	1.6 7	1.4 1	1.2 7	1.4 3	1.3 5	1.6 1	1.5 5	1.3 3	1.3 7	1.2 1	1.6 0	1.3 8
aɪ	1.4 3	1.3 5	1.4 7	1.3 5	1.6 2	1.4 4	1.2 6	1.4 5	1.7 0	1.4 3	1.3 8	1.4 4	1.4 8	1.3 0	1.2 2	1.6 4	1.3 3	1.3 8	1.5 3	1.3 6	1.6 7	1.3 7	1.6 1	1.4 9	1.4 4
ɔɪ	1.7 3	1.3 7	1.5 5	1.2 5	1.6 8	1.5 9	1.2 5	1.9 6	1.4 0	1.7 6	1.8 8	1.2 9	1.5 3	1.5 8	1.2 6	1.5 3	1.5 1	1.5 9	1.2 1	1.6 9	1.5 2	1.5 2	1.5 3	1.4 8	1.5 3
ɪə	1.3 1	1.0 0	1.3 1	1.3 8	1.7 3	1.1 7	1.2 9	1.5 9	1.1 5	1.3 4	1.4 1		1.1 6	0.8 2	1.2 2	1.2 7	1.6 3	1.5 5	1.5 8	1.3 1	1.3 1	1.6 9	1.5 3	1.3 1	1.3 5
eə	1.6 2	1.4 5	1.5 6	1.3 3	1.3 3	1.2 6		1.5 2	1.6 6	1.3 0	1.1 3	1.5 3	1.8 2	1.3 8	1.2 6	1.4 9	1.3 5	1.3 9	1.2 9	1.5 2	1.4 9	1.2 6	1.4 3	1.8 1	1.4 4
ʊə	1.5 3	1.7 1	1.2 0	1.3 1	2.0 0	1.3 7		1.4 2	1.2 2	1.6 9	1.7 6	1.4 0		1.6 1	1.5 3	1.1 9		1.8 1	0.9 2	1.3 6	1.2 2	1.4 3	1.5 5	1.0 8	1.4 4
ɪʊ	1.9 6	1.1 4	1.4 4	1.4 7	1.4 7	1.1 6	1.5 9	1.6 2	1.4 4	1.2 8	1.6 9	1.1 0	1.3 5	1.8 7	1.3 6	1.4 2	1.6 9	1.7 7	1.3 8	1.8 5	1.3 1	1.4 8	1.4 5	1.1 6	1.4 8
aʊ	1.8 9	1.5 6	1.2 8	1.2 6	1.2 2	1.3 1	1.4 4	1.4 0	1.0 6	1.6 0	1.5 3	1.2 5	1.4 5	1.4 2	1.7 8	1.3 9	1.6 1	1.3 5	1.5 9	1.5 4	1.2 7	1.5 0	1.6 5	1.6 0	1.4 6
əʊ	1.6 3	1.6 8	1.3 3	1.4 3	1.2 9	1.1 6	1.1 8	1.5 6	1.6 6	1.4 3	1.6 8	1.4 5	1.0 4	1.5 0	1.2 5	1.2 5	1.4 8	1.6 7	1.3 6	1.6 9	1.7 0	1.3 7	1.4 2	1.1 4	1.4 3
Me an	1.6 3	1.5 1	1.4 5	1.3 8	1.5 1	1.3 5	1.3 1	1.4 8	1.4 9	1.4 6	1.5 4	1.4 4	1.4 9	1.4 7	1.4 1	1.4 2	1.4 3	1.4 7	1.5 2	1.5 2	1.4 6	1.5 4	1.5 5	1.4 6	1.4 7

Table organisation is equivalent to that in Appendix 2 (see above)

## Appendix 2.4: Syllables Scores Minus Vowel Mean

	p	b	m	f	v	θ	ð	t	d	s	z	n	l	tʃ	dʒ	ʃ	ʒ	ɹ	j	k	g	ŋ	w	h
i	0.79	-	-	0.09	0.49	0.33	0.44	1.17	0.25	0.17	0.33	0.51	0.56	0.46	0.25	0.13	0.24	0.91	0.76	1.33	0.53	1.15	0.63	1.13
ɪ	0.07	0.86	0.53	0.16	0.23	0.45	0.25	1.15	0.21	0.09	0.36	0.41	0.08	0.40	0.01	0.13	0.77	0.41	0.79	1.18	0.47	0.17	0.97	0.25
e	0.05	0.33	0.37	0.09	0.35	0.07	0.05	0.67	0.01	0.08	0.59	0.33	0.33	0.79	0.35	0.02	0.07	0.91	1.01	0.94	0.47	0.25	0.80	0.02
ɛ	0.31	0.08	0.30	0.32	0.55	0.11	0.97	0.67	0.13	0.13	0.92	0.29	0.26	0.91	0.18	0.03	0.03	0.52	0.29	0.41	0.63	0.53	0.89	0.49
æ	0.15	0.33	0.52	0.05	0.43	0.09	0.27	0.71	0.04	0.55	0.87	0.91	0.81	0.48	0.19	0.37	0.01	0.33	0.64	1.47	0.13	0.09	0.87	0.11
ɜ	0.80	0.26	1.15	0.20	0.19	0.56	0.26	0.74	0.08	0.07	0.56	0.33	0.26	1.28	0.22	0.36	0.16	0.76	0.84	0.72	0.63	0.60	0.88	0.19
ə	0.36	0.18	0.53	0.68	0.18	0.56	0.25	0.98	0.23	0.18	0.77	0.34	0.38	0.86	0.70	0.22	0.06	0.71	0.38	1.10	0.02	0.79	0.78	0.46
ʌ	0.47	0.13	0.65	0.09	0.07	0.37	0.49	1.19	0.17	0.15	0.26	0.19	1.04	0.33	0.48	0.37	0.19		0.53	1.19	0.59	0.45	1.04	0.22
ɑ	0.13	0.29	0.93	0.17	0.20	0.65	0.01	0.79	0.09	0.13	0.55	0.25	0.55	0.95	0.07	0.17	0.90	0.55	0.01	1.95	0.13	0.01	1.21	0.21
ɑ:	0.46	0.40	0.54	0.26	0.58	0.50	0.42	0.74	0.26	0.26	0.54	0.14	0.94	0.54	0.34	0.22	0.52	0.18	0.70	1.42	0.50	0.74	1.30	1.04
ɒ	0.53	0.25	0.35	0.20	0.27	0.23	0.31	0.59	0.16	0.17	0.65	0.28	0.76	0.95	0.55	0.32	0.10		0.50	0.95	0.43	0.17	1.09	0.61
ɔ	0.62	0.20	0.68	0.06	0.32	0.10	0.10	1.18	0.65	0.26	0.26	0.39	0.38	0.76	0.58	0.10	0.50	0.24	0.94	0.73	0.74	0.65	1.13	0.22
ɔ:	0.06	0.33	0.30	0.10	0.11	0.34	0.90	0.78	0.26	0.14	0.81	0.45	0.72	1.14	0.33	0.22	0.62	0.70	0.46	1.12		0.45	1.30	0.50
u	0.08	0.15	1.32	0.00	0.11	0.55		1.24	0.21	0.22	1.11	0.41	0.22	0.85	0.85	0.41	0.36	0.60	0.19	1.20	0.11	0.78	0.88	1.52
ʊ	0.59	0.29	0.73	0.13	0.08	0.56	0.44	1.11	0.11	0.17	0.25	0.33	0.58	0.56	0.67	0.07	0.08		0.45	1.35	0.23	0.33	1.84	0.08
eɪ	0.56	0.28	0.48	0.15	0.15	0.04	0.06	0.15	0.00	0.15	0.04	0.34	0.41	0.15	0.64	0.12	0.04	0.57	0.76	1.48	0.34	0.28	1.00	0.03

aɪ	0.65	-	-	0.07	0.40	0.27	0.07	0.44	0.17	0.07	0.45	0.27	0.43	0.52	0.61	0.19	0.27	0.34	0.71	1.29	0.23	0.47	0.45	0.35		
ɔɪ	0.19	-	-	0.26	0.37	0.15	0.11	0.52	0.15	0.07	0.70	0.07	0.47	0.89	0.41	0.07	0.54	0.04	0.74	0.74	0.49	0.31	0.89	0.56		
ɪə	0.05	-	-	0.37	0.09	0.02	0.46	0.43	0.17	0.09	0.01	0.83	-	0.47	1.03	0.75	0.41	0.06	0.37	0.74	1.11	0.32	0.29	0.57	0.05	
eə	0.15	-	-	0.41	0.23	0.41	0.51	0.05	-	0.22	0.59	0.09	0.91	0.19	0.56	0.74	0.16	0.04	0.03	0.41	0.82	1.11	0.33	0.41	0.70	0.16
ʊə	0.08	-	-	0.06	0.92	0.36	0.32	0.10	-	0.13	0.09	0.39	0.91	0.52	-	0.72	0.36	0.48	0.06	0.87	0.99	0.68	0.96	0.84	0.20	
ɪʊ	0.09	-	-	0.47	0.58	0.65	0.08	0.19	0.41	0.38	0.35	0.45	1.15	0.14	0.36	0.97	0.19	0.05	0.19	0.33	0.37	0.68	0.60	0.07	0.81	0.85
aʊ	0.15	-	-	0.51	0.31	0.47	0.35	0.27	0.59	0.90	0.25	0.46	0.78	0.51	0.40	0.53	0.49	0.34	0.75	0.56	0.28	1.05	0.84	0.36	0.95	0.21
əʊ	0.38	-	-	0.43	0.77	0.31	0.01	0.09	0.16	0.50	0.01	0.09	1.03	0.17	1.05	0.15	0.71	0.03	0.17	0.37	0.89	0.72	0.51	0.37	0.76	0.49
Mean	0.17	-	-	0.23	0.59	0.06	0.17	0.01	0.02	0.71	0.07	0.02	0.63	0.33	0.52	0.71	0.40	0.01	0.13	0.46	0.61	1.09	0.37	0.42	0.94	0.36

*Syllable scores minus the mean syllable score for the syllable's vowel. Otherwise, the table is arranged as Appendix 2 (above). Intuitively put, these scores represent how much each consonant raised or lowered the baseline iconicity for each vowel, thus reading down columns shows how consistently consonants contributed roundness vs. spikiness to their syllable.*

## Appendix 2.5: Syllables Scores Minus Consonant Mean

	p	b	m	f	v	θ	ð	t	d	s	z	n	l	tʃ	dʒ	ʃ	ʒ	ʃ	j	k	g	ŋ	w	h	Mean
i	1.11	0.37	0.85	0.64	0.81	0.82	0.87	0.95	0.67	0.34	0.19	0.32	0.44	0.24	0.16	0.64	0.58	0.08	0.34	0.73	0.62	0.24	0.81	0.29	0.49
ɪ	0.34	0.19	0.49	0.34	0.03	0.02	0.63	0.87	0.58	0.55	0.17	0.36	0.86	0.13	0.03	0.32	1.06	0.52	0.26	0.52	0.50	0.68	0.41	1.04	0.44
e	0.11	0.13	0.45	0.37	0.40	0.30	0.13	0.19	0.17	0.11	0.19	0.24	0.41	0.32	0.18	0.22	0.16	0.19	0.17	0.08	0.30	0.40	0.37	0.56	0.23
ɛ	0.54	0.55	0.68	0.77	0.77	0.27	0.63	0.35	0.19	0.29	0.69	0.44	0.64	0.60	0.18	0.44	0.28	0.37	0.71	0.29	0.62	0.28	0.45	0.26	0.39
æ	0.26	0.18	0.35	0.28	0.53	0.18	0.49	0.27	0.24	0.25	0.51	0.29	0.03	0.05	0.07	0.66	0.12	0.44	0.25	0.65	0.00	0.60	0.35	0.75	0.28
ɜ	0.70	0.03	0.50	0.32	0.30	0.62	0.26	0.09	0.07	0.16	0.01	0.06	0.31	0.64	0.12	0.28	0.08	0.20	0.17	0.31	0.28	0.12	0.13	0.23	0.06
ə	0.56	0.77	0.42	0.26	0.37	0.21	0.56	0.63	0.06	0.56	0.50	0.36	0.49	0.52	0.66	0.60	0.28	0.14	0.59	0.37	0.06	0.00	0.53	0.26	0.36
ʌ	0.70	0.49	0.33	0.36	0.29	0.75	0.15	0.87	0.49	0.27	0.02	0.54	0.14	0.02	0.47	0.04	0.06		0.47	0.49	0.58	0.37	0.30	0.97	0.37
ɑ	0.02	0.21	0.07	0.15	0.29	0.38	0.21	0.35	0.11	0.17	0.19	0.36	0.23	0.52	0.06	0.12	1.02	0.22	0.87	1.13	0.26	0.70	0.00	0.42	0.27
ɑ:	0.06	0.53	0.31	0.03	0.05	0.86	0.01	0.33	0.17	0.58	0.45	0.16	0.79	0.52	0.42	0.12	0.02	0.04	0.45	0.03	0.26	0.68	0.71	1.03	0.35
ɒ	0.65	0.05	0.31	0.00	0.17	0.30	0.33	0.00	0.11	0.27	0.09	0.13	0.11	0.32	0.22	0.41	0.02		0.19	0.00	0.10	0.32	0.00	0.11	0.05

	3			7			0	5	6				7						7			7	8		
ɔ	0.8 2	0.0 1	0.1 3	0.0 8	0.1 1	0.0 6	0.1 9	0.4 3	0.5 4	0.2 7	0.4 1	0.0 9	0.0 9	0.0 2	0.1 4	0.1 2	0.3 2	0.2 2	0.3 7	0.4 0	0.3 0	0.2 6	0.2 2	0.1 0	0.0 4
ɔ:	0.9 0	0.9 0	0.5 0	0.6 4	0.8 5	0.4 6	0.0 5	0.7 3	0.6 1	0.6 3	0.6 1	0.9 0	1.0 1	0.3 6	0.8 6	1.0 0	1.5 6	1.0 0	0.6 5	0.7 7		0.8 2	1.1 5	0.9 4	0.7 7
u	0.7 4	0.4 1	1.2 3	0.4 4	0.7 8	0.0 5		0.0 3	0.7 8	0.6 9	0.0 2	0.5 7	0.2 1	0.3 5	0.0 4	0.0 8	0.2 8	0.6 0	0.0 7	0.3 9	0.7 9	0.8 6	0.4 3	1.6 6	0.4 9
ʊ	0.6 6	0.1 8	0.0 9	0.1 6	0.0 2	0.7 9	0.2 6	0.6 3	0.2 7	0.4 3	0.6 5	0.2 4	0.1 6	0.0 8	0.5 0	0.1 8	0.1 7		0.3 9	0.4 9	0.0 6	0.3 2	0.6 7	0.6 7	0.2 1
eɪ	0.3 8	0.4 9	0.0 9	0.1 9	0.3 4	0.0 6	0.0 1	0.5 8	0.0 9	0.1 6	0.6 1	0.0 1	0.0 9	0.5 7	0.2 2	0.1 2	0.1 2	0.0 9	0.1 7	0.3 7	0.7 5	0.6 8	0.0 7	0.3 8	0.0 1
aɪ	0.6 6	0.2 9	0.1 3	0.3 0	0.4 0	0.1 0	0.1 9	0.1 0	0.2 7	0.2 7	0.0 1	0.2 4	0.2 5	0.0 1	0.3 8	0.0 0	0.2 4	0.3 3	0.0 8	0.3 7	0.4 6	0.1 2	0.6 7	0.1 8	0.1 8
ɔɪ	0.3 3	0.3 8	0.0 7	0.5 5	0.1 5	0.2 1	0.5 2	0.5 4	0.5 7	0.4 0	0.2 8	0.0 6	0.3 1	0.1 6	0.3 4	0.2 6	1.0 4	0.1 8	0.4 8	0.7 0	0.2 6	0.2 4	0.3 0	0.5 5	0.3 5
ɪə	0.0 7	0.0 1	0.0 2	0.1 1	0.0 0	0.3 2	0.5 3	0.4 0	0.0 1	0.1 7	0.3 5		0.1 8	0.4 8	0.5 0	0.2 4	0.0 6	0.2 8	0.0 2	0.1 7	0.0 6	0.2 8	0.5 2	0.4 5	0.1 4
eə	0.3 7	0.2 3	0.3 1	0.4 0	0.2 9	0.1 0		0.5 4	0.4 7	0.1 1	0.2 3	0.1 0	0.1 0	0.0 2	0.2 9	0.0 8	0.1 6	0.0 4	0.2 6	0.0 3	0.1 2	0.0 4	0.1 9	0.1 5	0.0 5
ʊə	0.5 8	0.1 6	0.6 7	0.0 8	0.1 9	0.2 5		0.9 1	0.5 0	0.0 8	0.0 6	0.5 2		0.3 2	0.3 8	0.8 0		0.2 2	0.5 9	0.4 4	0.0 6	0.8 8	0.2 3	0.1 8	0.3 5
ɪʊ	0.9 8	0.9 7	0.7 2	1.3 2	0.8 2	0.5 4	0.3 7	1.0 6	0.4 5	1.1 5	0.2 1	0.5 3	0.5 8	0.4 6	0.9 4	0.7 6	0.6 8	0.5 6	0.4 9	1.1 5	0.5 3	0.3 7	0.5 9	1.2 2	0.7 3
aʊ	0.3 3	0.3 3	0.2 5	0.4 4	0.5 5	0.3 3	0.6 6	0.1 6	0.3 3	0.4 5	0.1 2	0.2 2	0.0 9	0.2 2	0.0 7	0.3 3	0.5 8	0.0 0	0.3 0	0.0 0	0.4 1	0.0 3	0.0 0	0.1 2	0.0 0

	4	1		4	5	0	7		5			0		0				9		7			4		3	
	-		-		-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
əʊ	0.1	0.2	0.5	0.0	0.5	0.2	0.2	0.5	0.4	0.2	0.0	0.2	0.9	0.9	0.0	0.3	0.6	0.2	0.6	0.7	0.2	0.3	0.1	0.5	0.3	
	5	9	5	0	5	8	6	8	5	5	3	0	1	2	6	2	8	4	5	4	6	2	9	0	6	

*Syllable scores minus the mean syllable score for the syllable's consonant. Otherwise, the table is arranged as Appendix 2 (above). Intuitively put, these scores represent how much each vowel raised or lowered the baseline iconicity for each vowel, thus reading across rows shows how consistently vowel contributed roundness vs. spikiness to their syllable.*

## Appendix 2.6: Specification for the Stops Mixed Model

Syllable\_score ~ (Stop\_lab + Stop\_av)\*(Stop\_nasal + Stop\_voiced)\*(Back.1((Stop\_lab + Stop\_av) + (Stop\_nasal + Stop\_voiced)|Vowel\_no)

## Appendix 2.7: Specification for the Fricatives Mixed Model

Mean ~ (Labial.1 + Dental + Alveolar)\*Voice\*Back.1 + ((Labial.1 + Dental + Alveolar) + Voice |Vowel\_no)

## Appendix 3.1: Ratings Scale for LetterScore Norming

na	<input type="radio"/>	1	2	3	4	5	6	7	8	9	10	<input type="radio"/>
va	<input type="radio"/>	1	2	3	4	5	6	7	8	9	10	<input type="radio"/>
ka	<input type="radio"/>	1	2	3	4	5	6	7	8	9	10	<input type="radio"/>

## Appendix 3.2: LetterScore Syllables and their Ratings

*LetterScore syllables and their ratings, arranged spikiest to roundest in descending order*

Syllable	MeanRating	CentredRating
ki	7.39	2.35
ke	7.25	2.21
ka	7.07	2.03
ze	6.96	1.92
zi	6.82	1.78
ti	6.79	1.75

vi	6.71	1.67
te	6.71	1.67
yi	6.64	1.60
ji	6.57	1.53
za	6.39	1.35
fi	6.21	1.17
pi	6.08	1.04
ni	6.07	1.03
ta	6.04	1.00
di	6.00	0.96
ku	5.93	0.89
pe	5.89	0.85
li	5.86	0.82
zu	5.79	0.75
ri	5.75	0.71
de	5.71	0.67
ve	5.71	0.67
fe	5.67	0.63
re	5.61	0.57
je	5.61	0.57
ko	5.57	0.53
tu	5.54	0.50
ya	5.54	0.50
si	5.50	0.46
ja	5.43	0.39
va	5.39	0.35
ye	5.32	0.28
vu	5.32	0.28
se	5.29	0.25
wa	5.29	0.25
zo	5.25	0.21
bi	5.25	0.21
we	5.21	0.17
ha	5.14	0.10
wi	5.14	0.10
pu	5.11	0.07
ne	5.11	0.07
fa	5.11	0.07
sa	5.04	0.00
hi	5.04	0.00
mi	4.89	-0.15

to	4.86	-0.18
na	4.86	-0.18
ra	4.71	-0.33
ru	4.68	-0.36
da	4.64	-0.40
nu	4.61	-0.43
he	4.61	-0.43
be	4.57	-0.47
fu	4.54	-0.50
vo	4.50	-0.54
pa	4.43	-0.61
le	4.36	-0.68
ba	4.32	-0.72
jo	4.32	-0.72
po	4.29	-0.75
ro	4.25	-0.79
fo	4.21	-0.83
no	4.21	-0.83
la	4.21	-0.83
yu	4.19	-0.85
me	4.18	-0.86
hu	4.18	-0.86
bu	4.14	-0.90
ju	3.96	-1.08
du	3.96	-1.08
su	3.96	-1.08
wu	3.86	-1.18
do	3.64	-1.40
so	3.57	-1.47
lu	3.54	-1.50
yo	3.50	-1.54
ma	3.50	-1.54
bo	3.46	-1.58
wo	3.46	-1.58
ho	3.43	-1.61
lo	3.29	-1.75
mu	3.18	-1.86
mo	2.82	-2.22

## Appendix 3.3: Specifications for Mixed Model for Experiment 3.1

*The first successfully converged omnibus model for Experiment 3.1*

Correct ~ poly(Block,2,raw=T)\*Congruent\*Diff\_Cat + (Block + Congruent\*Diff\_Cat|Subject\_Code) + (Congruent\*Diff\_Cat|Name)

## Appendix 3.4: Specifications for Mixed Model for Experiment 3.2

*The first successfully converged omnibus model used to analyse Experiment 3.2*

Correct ~ poly(Block,2,raw=T)\*Condition\_Type\*Congruent\*Diff\_Cat - poly(Block,2,raw=T):Condition\_Type:Congruent:Diff\_Cat + (poly(Block,2,raw=T)|Subject\_Code) + (Condition\_Type - 1 |Subject\_Code) + (Congruent + Diff\_Cat - 1|Subject\_Code) + (poly(Block,2,raw=T)|Name) + (Diff\_Cat - 1|Name)

I was forced to restrict the covariance structure so as to arrive at a model that would converge. Random effects terms within a given set of brackets had covariance parameters estimated within the model. We chose these bracketings according to principled criteria. Thus we allowed the intercept and the block slope to covary for both subjects and names, as we expected the two to be correlated, as both will reflect rate of learning. Moreover we allowed congruence and category of foil to covary as both are implicated in the congruence\*foil interaction we found in Experiment 1 (a subject with strong sensitivity to congruence may have a steep slope for both).

## Appendix 3.5: Specifications for Mixed Model for Experiment 3.3

*The omnibus model for Experiment 3.3*

Correct ~ poly(Block,2,raw=T)\*Condition\_Type\*Congruent\*Diff\_Cat -  
poly(Block,2,raw=T):Condition\_Type:Congruent:Diff\_Cat +  
(poly(Block,2,raw=T)|Subject\_Code) + (Condition\_Type - 1 |Subject\_Code) +  
(Congruent + Diff\_Cat - 1|Subject\_Code) + (poly(Block,2,raw=T)|Name) +  
(Congruent + Diff\_Cat - 1|Name)

## Appendix 3.6: Specifications for Mixed Model for Experiment 3.4

*The omnibus model for Experiment 3.4*

Correct ~ poly(Block,2,raw=T)\*Condition\_Type\*Congruent\*Diff\_Cat -  
poly(Block,2,raw=T):Condition\_Type:Congruent:Diff\_Cat +  
(poly(Block,2,raw=T)|Subject\_Code) + (Congruent + Diff\_Cat - 1|Subject\_Code) +  
(poly(Block,2,raw=T)|Name) + (Congruent + Diff\_Cat - 1|Name)

## Appendix 3.7: Specifications for Bayesian Models for Chapter 3

*Bayesian model specification for overall analysis of Experiments 3.2-3.4*

Correct ~ (Block\_Normalised +  
Block\_Squared)\*Condition\_Type\*Congruent\*Diff\_Cat\*Condition\_Ordinal\_Centred -  
(Block\_Normalised +  
Block\_Squared):Condition\_Type:Congruent:Diff\_Cat:Condition\_Ordinal\_Centred -  
(Block\_Normalised + Block\_Squared):Condition\_Type:Congruent:Diff\_Cat -  
(Block\_Normalised +  
Block\_Squared):Condition\_Type:Congruent:Condition\_Ordinal\_Centred -  
(Block\_Normalised +  
Block\_Squared):Condition\_Type:Diff\_Cat:Condition\_Ordinal\_Centred -  
(Block\_Normalised +  
Block\_Squared):Congruent:Diff\_Cat:Condition\_Ordinal\_Centred -  
Condition\_Type:Congruent:Diff\_Cat:Condition\_Ordinal\_Centred+  
((Block\_Normalised + Block\_Squared)|Subject\_Code) + (Congruent + Diff\_Cat -  
1|Subject\_Code) + (Condition\_Ordinal\_Centred - 1|Subject\_Code) +  
((Block\_Normalised + Block\_Squared)|Name) + (Congruent + Diff\_Cat - 1|Name)

## Appendix 3.8: 95% Highest Density Intervals for Bayesian Models for Chapter 3

*95% highest density intervals for Bayesian Model 1*

Parameter	Mean	2.50%	97.50%	Do HDIs exclude zero?
(Intercept)	1.516	1.372	1.662	1
Block_Normalised	1.498	1.350	1.653	1
Block_Squared	-0.105	-0.178	-0.030	1
Condition_Type	-0.017	-0.164	0.127	0
Congruent	0.123	0.027	0.216	1
Diff_Cat	0.063	-0.003	0.128	0
Condition_Ordinal_Centred	0.495	0.360	0.633	1
Block_Normalised:Condition_Type	-0.091	-0.231	0.042	0
Block_Squared:Condition_Type	0.028	-0.080	0.140	0
Block_Normalised:Congruent	-0.101	-0.213	0.009	0
Block_Squared:Congruent	-0.035	-0.137	0.068	0
Condition_Type:Congruent	0.006	-0.122	0.139	0
Block_Normalised:Diff_Cat	-0.084	-0.192	0.021	0
Block_Squared:Diff_Cat	0.002	-0.099	0.105	0
Condition_Type:Diff_Cat	-0.060	-0.173	0.053	0
Congruent:Diff_Cat	0.326	0.220	0.431	1
Block_Normalised:Condition_Ordinal_Centred	0.324	0.192	0.452	1
Block_Squared:Condition_Ordinal_Centred	-0.202	-0.315	-0.091	1
Condition_Type:Condition_Ordinal_Centred	0.197	-0.385	0.799	0
Congruent:Condition_Ordinal_Centred	-0.008	-0.123	0.107	0
Diff_Cat:Condition_Ordinal_Centred	-0.062	-0.173	0.050	0
Block_Normalised:Condition_Type:Congruent	0.187	-0.023	0.391	0
Block_Squared:Condition_Type:Congruent	-0.012	-0.222	0.193	0
Block_Normalised:Condition_Type:Diff_Cat	0.064	-0.148	0.276	0
Block_Squared:Condition_Type:Diff_Cat	0.052	-0.156	0.269	0
Block_Normalised:Congruent:Diff_Cat	-0.072	-0.282	0.137	0
Block_Squared:Congruent:Diff_Cat	-0.029	-0.221	0.168	0
Condition_Type:Congruent:	-0.089	-0.284	0.098	0

Diff_Cat				
Block_Normalised:Condition_Type:Condition_Ordinal_Centred	-0.219	-0.807	0.393	0
Block_Squared:Condition_Type:Condition_Ordinal_Centred	-0.134	-0.416	0.159	0
Block_Normalised:Congruent:Condition_Ordinal_Centred	-0.005	-0.227	0.223	0
Block_Squared:Congruent:Condition_Ordinal_Centred	-0.222	-0.436	-0.007	1
Condition_Type:Congruent:Condition_Ordinal_Centred	0.010	-0.281	0.321	0
Block_Normalised:Diff_Cat:Condition_Ordinal_Centred	0.049	-0.170	0.260	0
Block_Squared:Diff_Cat:Condition_Ordinal_Centred	-0.070	-0.279	0.141	0
Condition_Type:Diff_Cat:Condition_Ordinal_Centred	-0.167	-0.393	0.059	0
Congruent:Diff_Cat:Condition_Ordinal_Centred	-0.271	-0.476	-0.059	1

*95% highest density intervals for Bayesian Model 2*

Parameter	Mean	2.50%	97.50%	Do HDIs exclude zero?
(Intercept)	1.714	1.519	1.922	1
Block_Normalised	1.638	1.450	1.835	1
Block_Squared	-0.155	-0.255	-0.060	1
Condition_Type	-0.013	-0.172	0.145	0
Congruent	0.087	-0.022	0.198	0
Diff_Cat	0.091	0.019	0.164	1
Condition_Ordinal	0.472	0.322	0.620	1
Block_Normalised:Condition_Type	-0.101	-0.239	0.043	0
Block_Squared:Condition_Type	0.028	-0.093	0.154	0
Block_Normalised:Congruent	-0.080	-0.213	0.053	0
Block_Squared:Congruent	-0.048	-0.176	0.075	0
Condition_Type:Congruent	-0.022	-0.165	0.125	0
Block_Normalised:Diff_Cat	-0.063	-0.191	0.066	0
Block_Squared:Diff_Cat	-0.037	-0.157	0.088	0
Condition_Type:Diff_Cat	-0.110	-0.241	0.021	0

Congruent:Diff_Cat	0.242	0.114	0.369	1
Block_Normalised:Condition_Ordinal	0.329	0.196	0.462	1
Block_Squared:Condition_Ordinal	-0.213	-0.333	-0.094	1
Condition_Type:Condition_Ordinal	0.301	-0.499	1.080	0
Congruent:Condition_Ordinal	0.004	-0.121	0.130	0
Diff_Cat:Condition_Ordinal	-0.112	-0.238	0.012	0
Block_Normalised:Condition_Type:Congruent	0.215	-0.035	0.469	0
Block_Squared:Condition_Type:Congruent	0.038	-0.204	0.272	0
Block_Normalised:Condition_Type:Diff_Cat	0.047	-0.207	0.297	0
Block_Squared:Condition_Type:Diff_Cat	0.124	-0.105	0.363	0
Block_Normalised:Congruent:Diff_Cat	-0.039	-0.292	0.209	0
Block_Squared:Congruent:Diff_Cat	0.043	-0.190	0.277	0
Condition_Type:Congruent:Diff_Cat	-0.112	-0.350	0.125	0
Block_Normalised:Condition_Type:Condition_Ordinal	-0.273	-1.034	0.491	0
Block_Squared:Condition_Type:Condition_Ordinal	-0.220	-0.572	0.124	0
Block_Normalised:Congruent:Condition_Ordinal	-0.050	-0.304	0.206	0
Block_Squared:Congruent:Condition_Ordinal	-0.253	-0.498	-0.009	1
Condition_Type:Congruent:Condition_Ordinal	0.130	-0.198	0.462	0
Block_Normalised:Diff_Cat:Condition_Ordinal	0.015	-0.237	0.272	0
Block_Squared:Diff_Cat:Condition_Ordinal	-0.011	-0.250	0.228	0
Condition_Type:Diff_Cat:Condition_Ordinal	-0.108	-0.360	0.159	0
Congruent:Diff_Cat:Condition_Ordinal	-0.162	-0.391	0.073	0

## Appendix 4.1: The Ratings Scale Used in the Norming Studies for WordScore in Experiments 4.1 and 4.2.



## Appendix 4.2: Final Model Specifications for Experiments 4.1 and 4.2.

### Experiment 4.1: Error

$$\text{poly}(\text{Ten\_Gen\_Centered}, 2, \text{raw}=\text{T}) * (\text{Shape\_Centered} + \text{Motion\_Linear} + (1|\text{CHAIN}/\text{GENERATION}) + (\text{poly}(\text{Ten\_Gen\_Centered}, 2, \text{raw}=\text{T}) - 1|\text{CHAIN}))$$

### Experiment 4.1: LetterScore

$$\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) * \text{Shape\_Centered} + \text{Motion\_Linear} + (\text{Shape\_Centered}|\text{CHAIN}/\text{GENERATION}) + (\text{poly}(\text{Eleven\_Gen\_Centered}, 1, \text{raw}=\text{T}) * \text{Shape\_Centered} - \text{Shape\_Centered} - 1|\text{CHAIN})$$

### Experiment 4.1: LetterScore for round/spiky stimuli only

$$\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) + (1|\text{CHAIN}/\text{GENERATION}) + (\text{poly}(\text{Ten\_Gen\_Centered}, 2, \text{raw}=\text{T}) - 1|\text{CHAIN})$$

### Experiment 4.1: WordScore

$\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) * (\text{Shape\_Centered} + \text{Motion\_Linear}) +$   
 $(\text{Shape\_Centered} | \text{CHAIN} / \text{GENERATION}) +$   
 $(\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) * \text{Shape\_Centered} - \text{Shape\_Centered} - 1 | \text{CHAIN})$

### Experiment 4.1: WordScore for round/spiky stimuli only

$\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) + (1 | \text{CHAIN} / \text{GENERATION}) +$   
 $(\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) - 1 | \text{CHAIN})$

### Experiment 4.1: Length in letters

$\text{poly}(\text{Eleven\_Gen\_Centered}, 1, \text{raw}=\text{T}) * \text{Motion\_Linear} + \text{Shape\_Centered} +$   
 $(\text{Motion\_Linear} | \text{CHAIN} / \text{GENERATION}) + (\text{poly}(\text{Ten\_Gen\_Centered}, 1, \text{raw}=\text{T}) * \text{Motion\_Linear} -$   
 $\text{Motion\_Linear} - 1 | \text{CHAIN})$

### Experiment 4.1: Length in letters for still/single motion/bouncing stimuli only

$\text{poly}(\text{Eleven\_Gen\_Centered}, 1, \text{raw}=\text{T}) + (1 | \text{CHAIN} / \text{GENERATION}) +$   
 $(\text{poly}(\text{Ten\_Gen\_Centered}, 2, \text{raw}=\text{T}) - 1 | \text{CHAIN})$

### Experiment 4.1: Length in syllables

$\text{poly}(\text{Eleven\_Gen\_Centered}, 1, \text{raw}=\text{T}) * (\text{Shape\_Centered} + \text{Motion\_Linear}) +$   
 $(\text{Motion\_Linear} | \text{CHAIN/GENERATION}) + (\text{poly}(\text{Ten\_Gen\_Centered}, 2, \text{raw}=\text{T}) * \text{Motion\_Linear} -$   
 $\text{Motion\_Linear} - 1 | \text{CHAIN})$

**Experiment 4.1: Length in letters for still/single motion/bouncing stimuli only**

$\text{poly}(\text{Eleven\_Gen\_Centered}, 1, \text{raw}=\text{T}) + (1 | \text{CHAIN/GENERATION}) +$   
 $(\text{poly}(\text{Ten\_Gen\_Centered}, 2, \text{raw}=\text{T}) - 1 | \text{CHAIN})$

**Experiment 4.2: WordScore**

$\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) * \text{Shape\_Centered} * \text{Motion\_Linear} +$   
 $(\text{Shape\_Centered} | \text{Chain/Generation}) +$   
 $(\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) * \text{Shape\_Centered} - \text{Shape\_Centered} - 1 | \text{Chain})$

**Experiment 4.2: WordScore for round/spiky stimuli only**

$\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) + (1 | \text{Chain/Generation}) +$   
 $(\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) - 1 | \text{Chain})$

**Experiment 4.1: Length in syllables**

$\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) * \text{Motion\_Linear} + \text{Shape\_Centered} +$   
 $(\text{Motion\_Linear} | \text{Chain/Generation}) + (\text{poly}(\text{Eleven\_Gen\_Centered}, 1, \text{raw}=\text{T}) * \text{Motion\_Linear} -$   
 $\text{Motion\_Linear} - 1 | \text{Chain})$

**Experiment 4.1: Length in letters for still/single motion/bouncing stimuli only**

$\text{poly}(\text{Eleven\_Gen\_Centered}, 2, \text{raw}=\text{T}) + (1 | \text{Chain/Generation}) +$   
 $(\text{poly}(\text{Eleven\_Gen\_Centered}, 1, \text{raw}=\text{T}) - 1 | \text{Chain})$

## Appendix 4.3: Analysis of Number of Syllables for Experiment 4.1

Syllables were coded by me assuming English phonotactics

Analysis of number of syllables yielded similar results to analysis of number of letters. The initial omnibus model shows reliable interactions between shape and linear generation, and manner of motion and linear generation. Other interactions, and quadratic generation, were not reliable ( $|t| < 1.9$ ), and were not included in the reduced model. In the reduced model, linear generation is a reliable predictor ( $\beta = -0.034$ , 95% CI [-0.063, -0.005],  $t = -2.30$ ), indicating an overall downwards trend in number of syllables, as is shape ( $\beta = -0.088$ , 95% CI [-0.166, -0.010],  $t = -2.21$ ), indicating longer names for rounded stimuli. Also reliable is the interaction between linear generation and shape ( $\beta = 0.053$ , 95% CI [0.028, 0.079],  $t = 4.12$ ) indicating that names for rounded stimuli lost syllables faster than names for spiky stimuli. Finally, the interaction between manner of motion and linear generation is reliable ( $\beta = 0.052$ , 95% CI [0.016, 0.089],  $t = 2.89$ ).

We fitted separate models for each manner of motion, featuring only linear generation as a predictor. Linear generation was a reliable effect for still ( $\beta = -0.091$ , 95% CI [-0.128, -0.053],  $t = -4.71$ ) stimuli, indicating an overall downwards trend in each case. However there was no reliable effect of linear generation for upwards moving stimuli ( $|t| < 1.9$ ), or bouncing stimuli ( $|t| < 1.0$ ). Thus in summary, the metric of number of syllables shows divergence between different manners of motion over the generations (much like the metric of length in letters). However, the divergence between different manners of motion is driven by still stimuli's names losing

syllables, in keeping with the overall downwards trend in number of syllables seen in the full model.

## Appendix 4.4: English Monolingual Analysis for Experiment 4.3.

**Participants**  $n = 24$ , twenty-two women,  $M = 18.8 \pm 0.8$ .

There was a significant difference between rounded ( $n = 9$ ,  $M = -0.454$ ,  $SD = 0.280$ , 95% CI of the mean = [-0.637, -0.271]) and spiky stimuli ( $n = 15$ ,  $M = 0.195$ ,  $SD = 0.403$ , 95% CI of the mean = [-0.009, 0.399]):  $t(22) = 4.24$ ,  $p < .001$ ; difference = 0.649, with 95% CI of the difference [0.331, 0.967], and Cohen's  $d = 1.807$ .

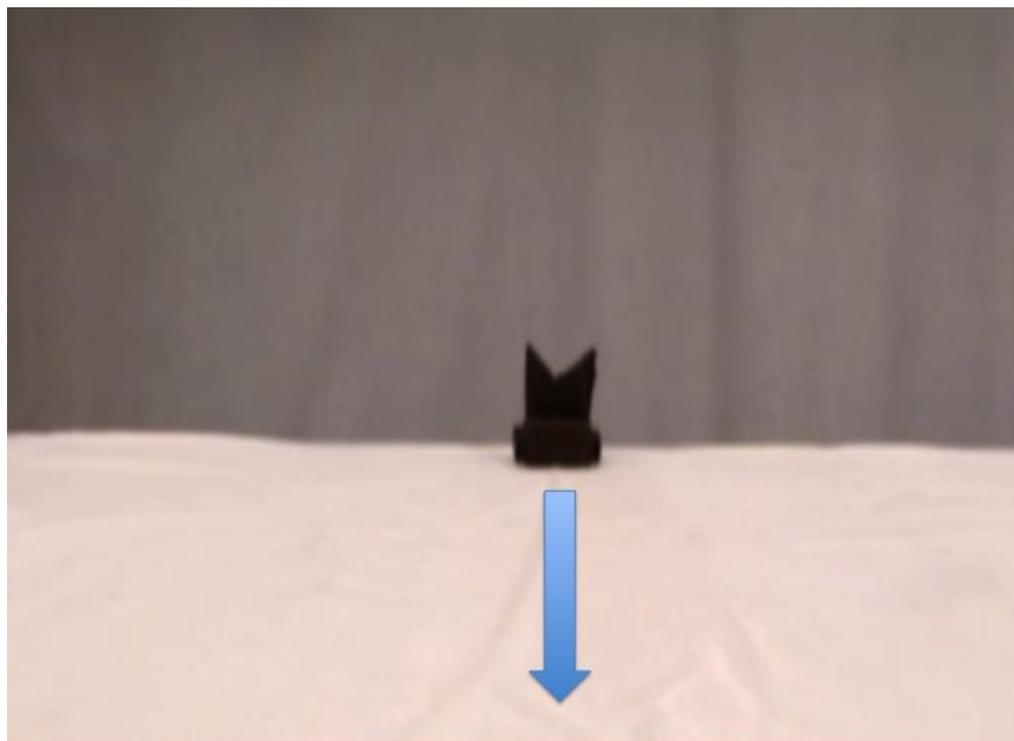
## Appendix 4.5: Syllables as a Motion-Iconicity Metric for Experiment 4.3

Syllables were coded assuming English phonotactics. Motion stimuli showed a significant difference (single motion median = 1, repeated motion median = 2;  $U = 222.0$ ,  $p = .025$ ).

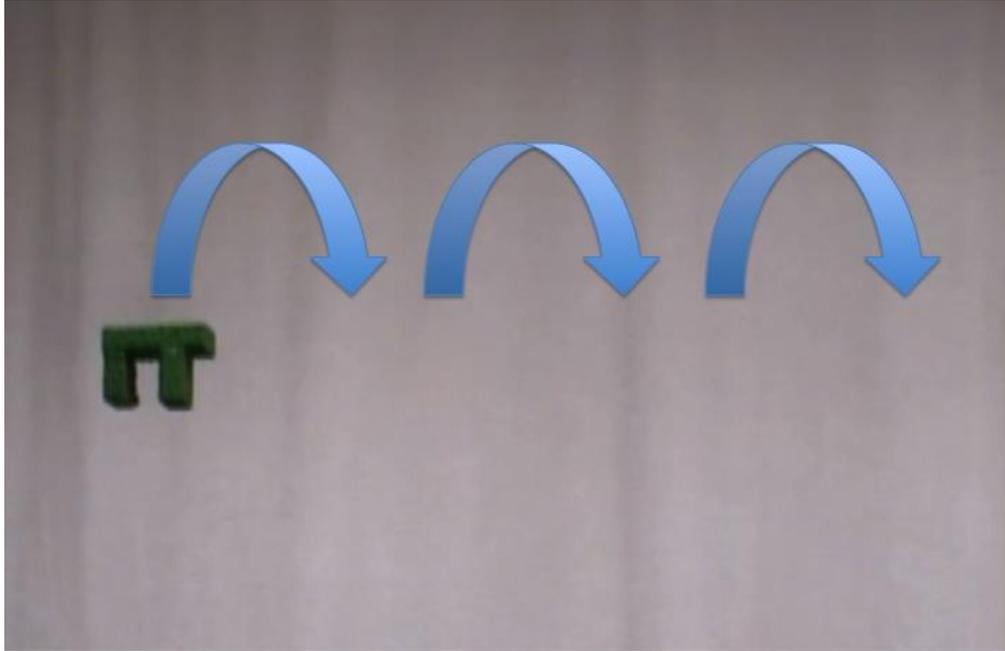
**English Monolingual Analysis** There was a significant difference in length between the condition, both in terms of number of letters (single median = 4, repeat median = 5;  $U = 28.5$ ,  $p = .010$ ) and number of syllables (single median = 1, repeat median = 2;  $U = 33.0$ ,  $p = .012$ ).

## Appendix 5.1: Object Videos

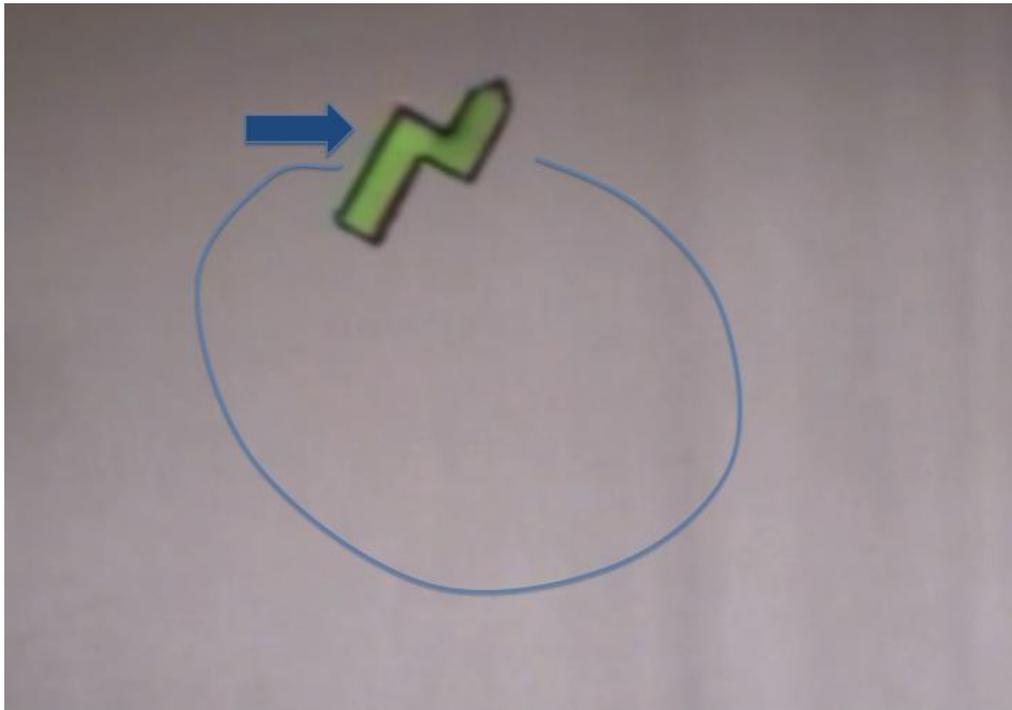
In each case blue arrows demonstrate the path of motion. All videos lasted 3.5 seconds.



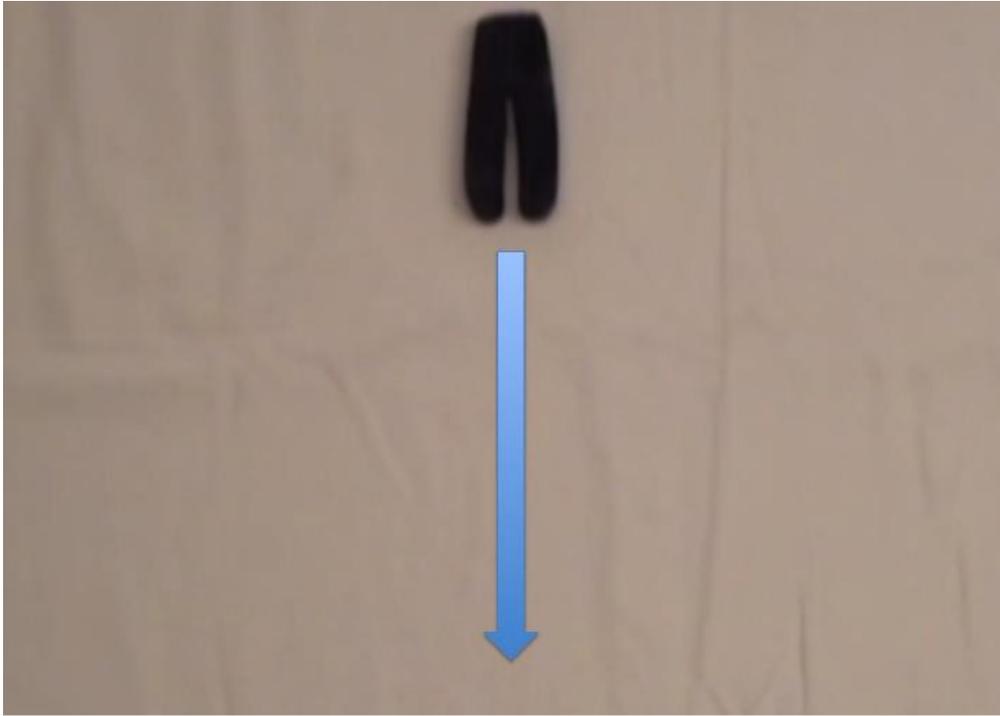
*Approach*



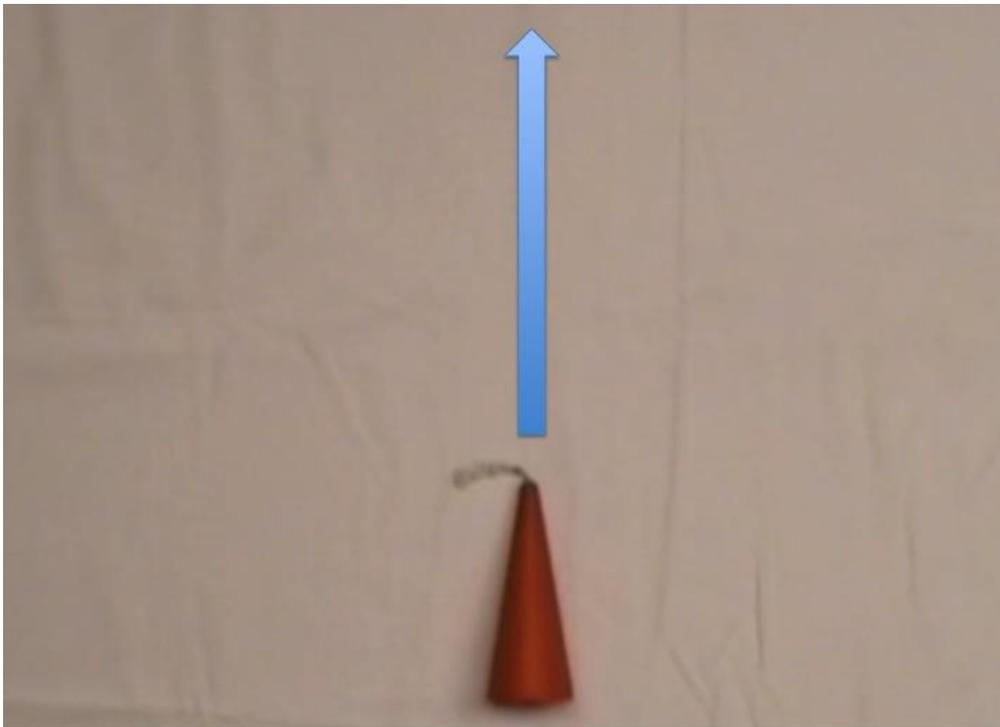
*Bounce*



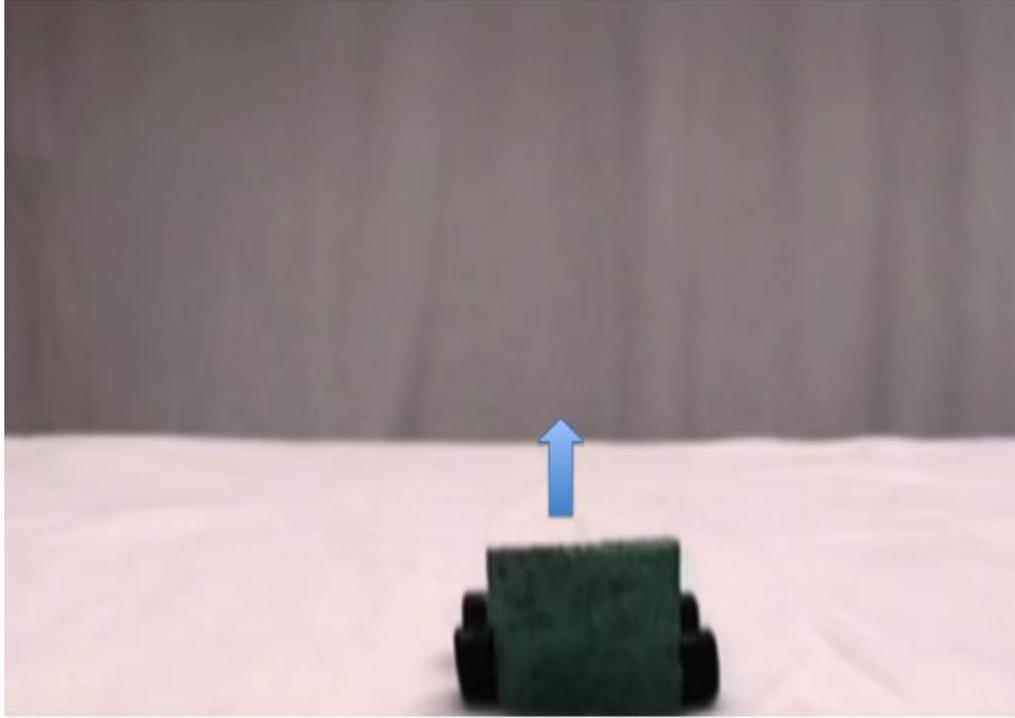
*Circle*



*Down*



*Up*



*Withdraw*

## Appendix 5.2: Gesture Videos

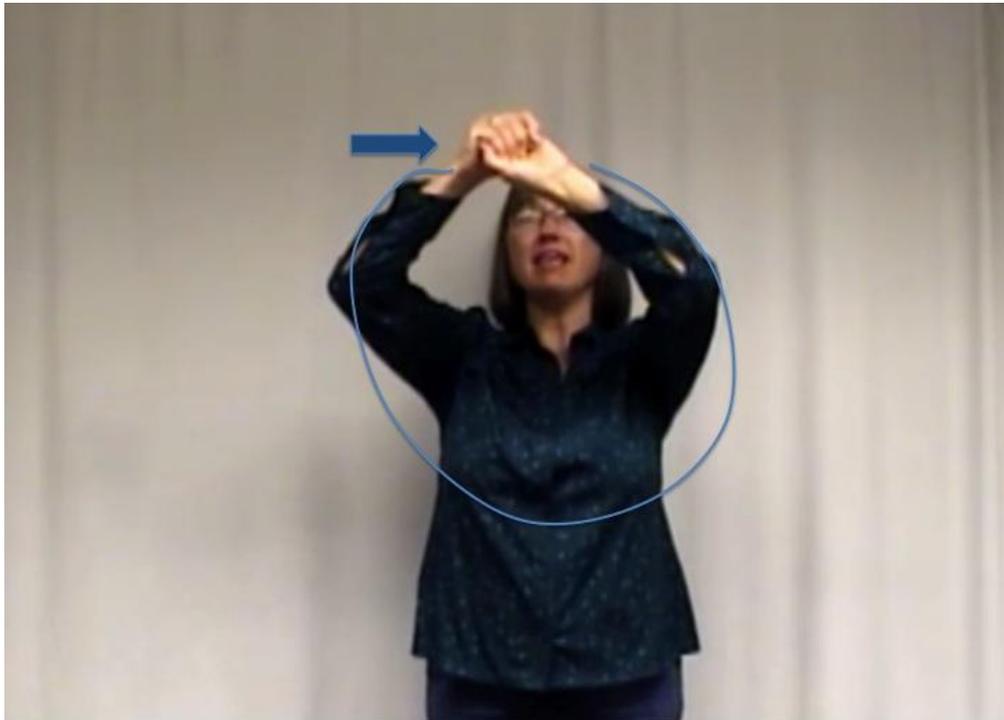
In each case blue arrows demonstrate the path of motion.



*Approach*



*Bounce*



*Circle*



*Down*



*Up*



Withdraw

## Appendix 5.3: Trial Orders

Two random orders were generated for the target object videos, each also having a random assignment of foils (within the constraint that each object video could only appear once as a target, and once as a foil). On top of this, each gesture video came in two versions (as discussed in the materials section), differing principally in the nonword the gesture was paired with. This resulted in  $2 \times 2 = 4$  separate versions of the gesture condition slideshow: two trial orders, each with both versions of the gesture videos.

Control condition slideshows were identical to gesture condition slideshows, except that the gesture videos were replaced with attention getters of equal duration. As slideshow 1 only differed from slideshow 2 in the content of its gesture video, these two slideshows were effectively the same for the control condition (ditto for slideshows 3 and 4).

Spatial arrangement of the target object video was counterbalanced such that in each slideshow it appeared on the left half of the time, and on the right the other half of the time.

*Slideshow 1*

<b>Trial Number</b>	<b>Target</b>	<b>Foil</b>	<b>Name</b>	<b>Target side</b>
1	Bounce	Withdraw	[pɛɪŋ]	Right
2	Up	Approach	[jɒfɪŋ]	Left
3	Approach	Circle	[gɑɪbɪŋ]	Left
4	Down	Bounce	[wæzɪŋ]	Right
5	Withdraw	Down	[dʒəʊpɪŋ]	Left
6	Circle	Up	[fɪmɪŋ]	Right

*Slideshow 2*

<b>Trial Number</b>	<b>Target</b>	<b>Foil</b>	<b>Name</b>	<b>Target side</b>
1	Bounce	Withdraw	[dʒəʊpɪŋ]	Right
2	Up	Approach	[wæzɪŋ]	Left
3	Approach	Circle	[fɪmɪŋ]	Left
4	Down	Bounce	[pɛɪŋ]	Right
5	Withdraw	Down	[gɑɪbɪŋ]	Left
6	Circle	Up	[jɒfɪŋ]	Right

*Slideshow 3*

<b>Trial Number</b>	<b>Target</b>	<b>Foil</b>	<b>Name</b>	<b>Target side</b>
1	Approach	Down	[gɑɪbɪŋ]	Left
2	Circle	Approach	[fɪmɪŋ]	Right
3	Down	Circle	[wæzɪŋ]	Right
4	Bounce	Up	[pɛɪŋ]	Right
5	Withdraw	Bounce	[dʒəʊpɪŋ]	Left
6	Up	Withdraw	[jɔfɪŋ]	Left

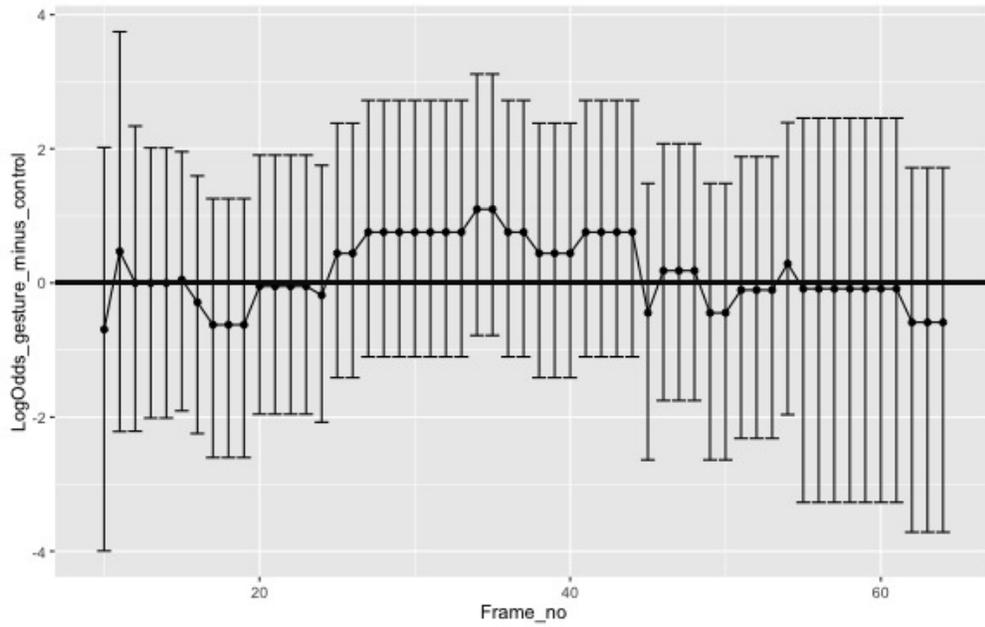
*Slideshow 4*

<b>Trial Number</b>	<b>Target</b>	<b>Foil</b>	<b>Name</b>	<b>Target side</b>
1	Approach	Down	[fɪmɪŋ]	Left
2	Circle	Approach	[jɔfɪŋ]	Right
3	Down	Circle	[pɛɪŋ]	Right
4	Bounce	Up	[dʒəʊpɪŋ]	Right
5	Withdraw	Bounce	[gɑɪbɪŋ]	Left
6	Up	Withdraw	[wæzɪŋ]	Left

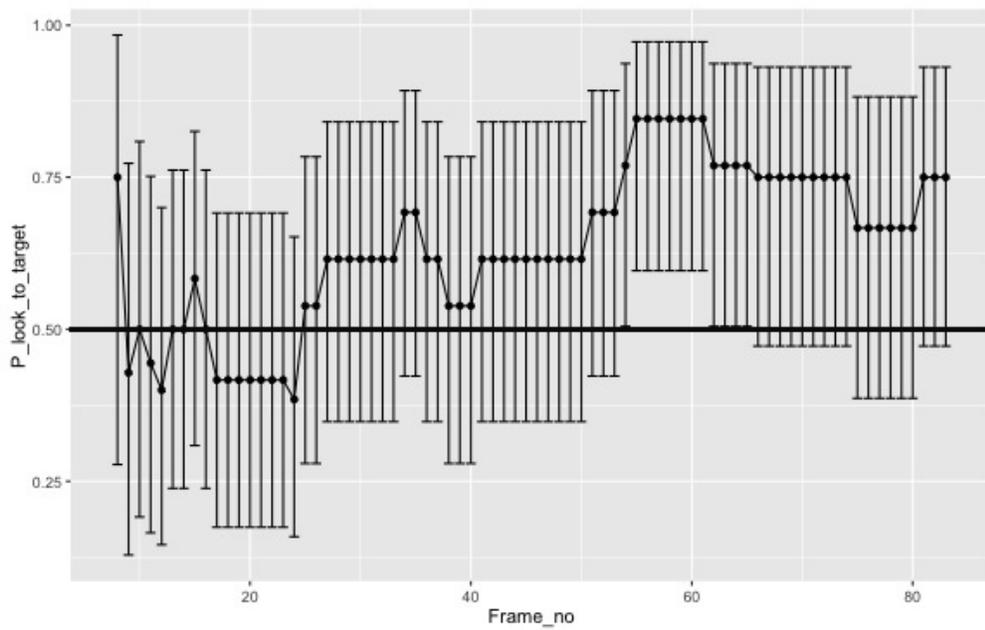
## Appendix 5.4: Analysis by items

All analyses here are of looks to target vs. foil by frame. Analyses were by binomial logistic regression (no random effects: each participant only had one trial with each item).

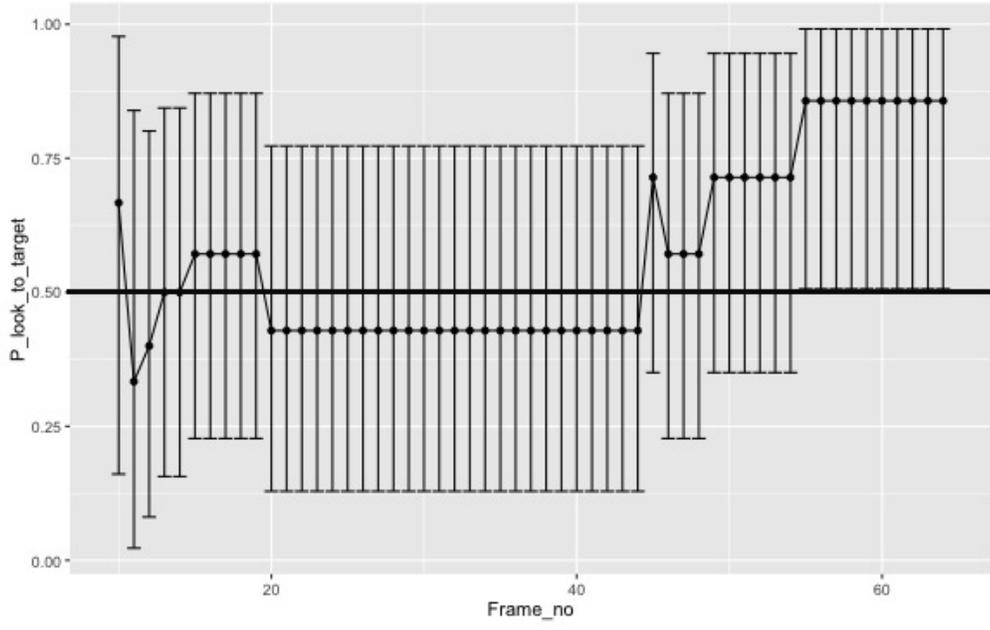
## Approach



*Gesture condition vs. control condition*

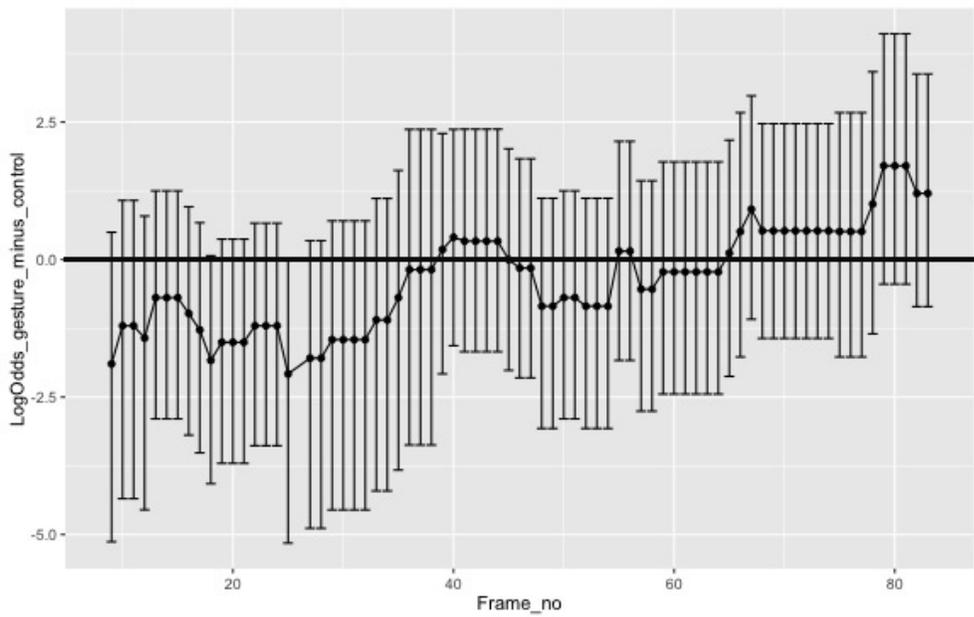


*Gesture condition*

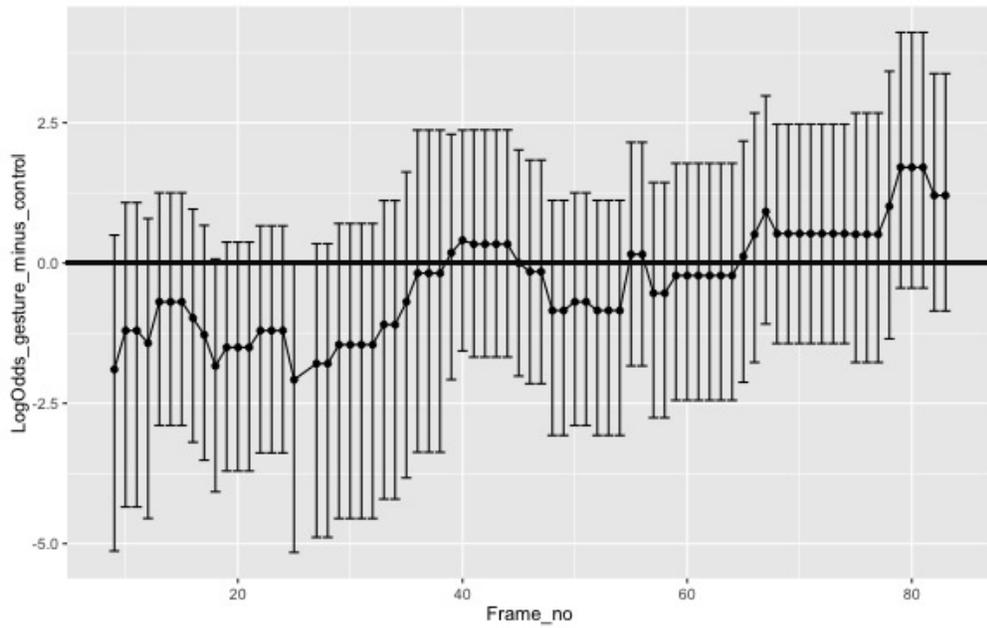


*Control condition*

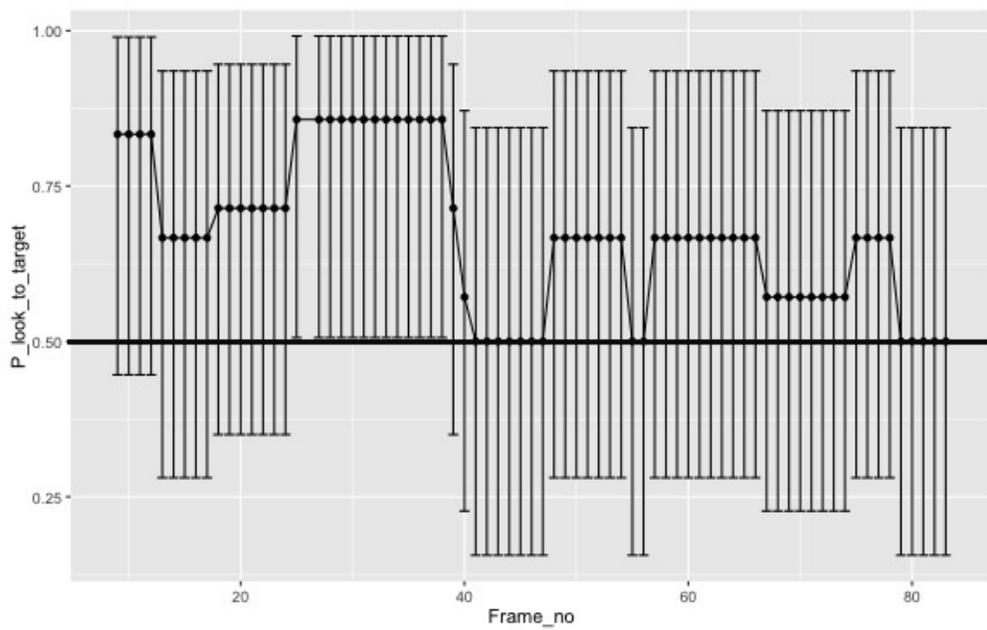
**Bounce**



*Gesture condition vs. control condition*

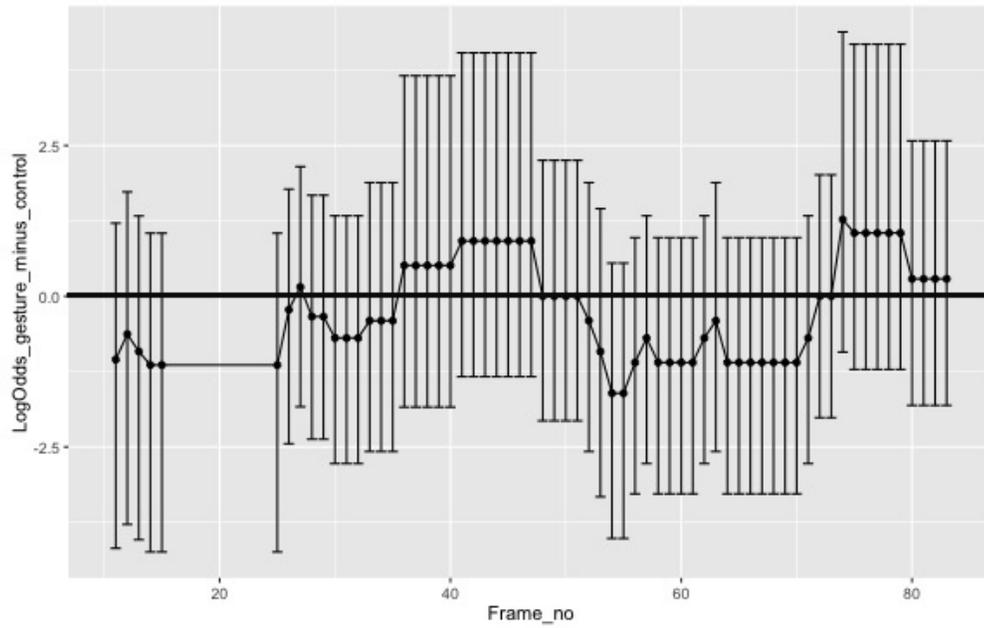


*Gesture condition*

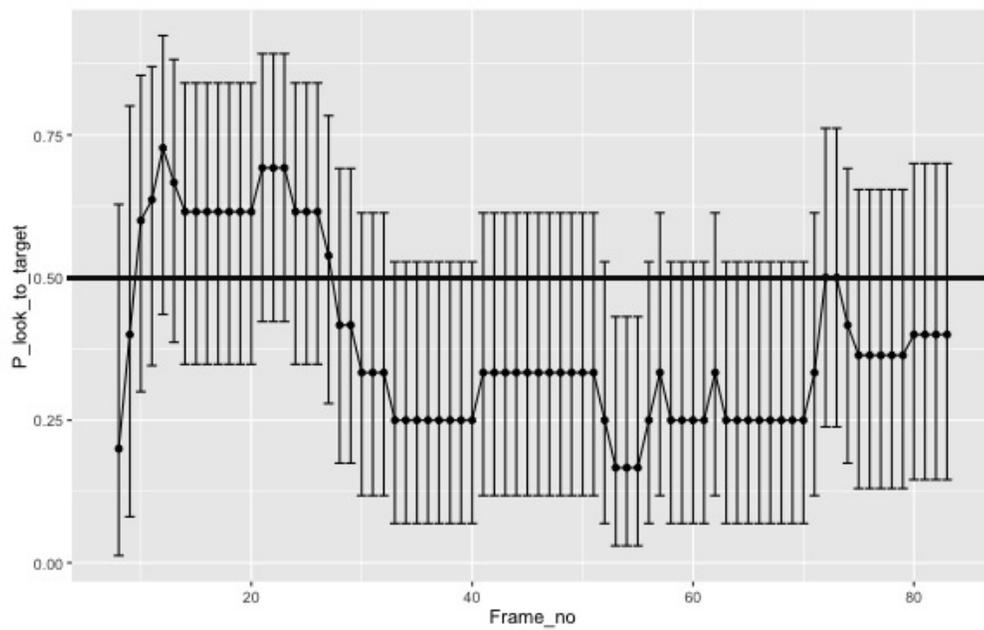


*Control Condition*

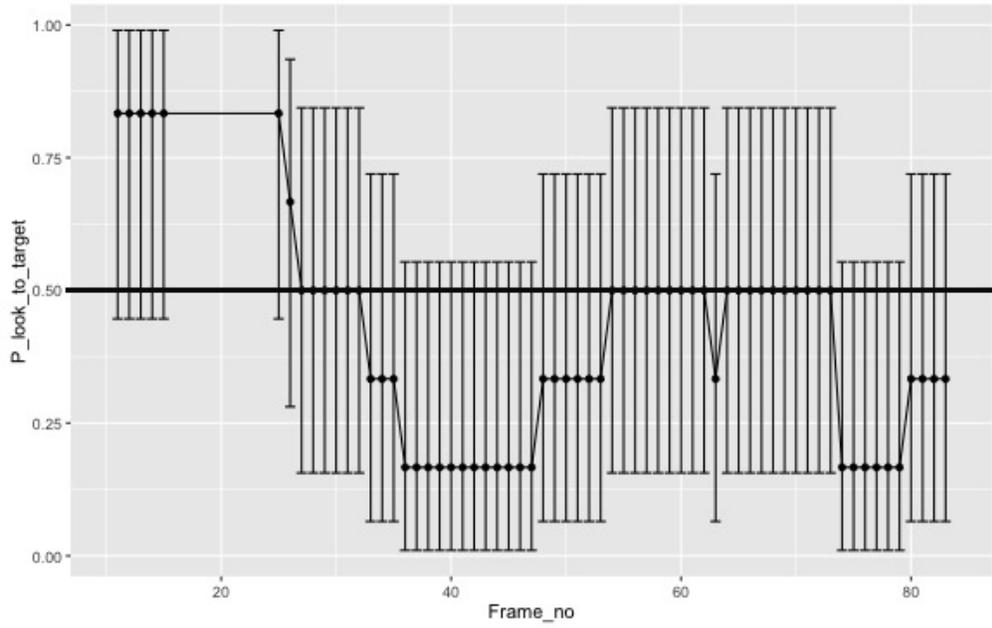
## Circle



*Gesture condition vs. control condition*

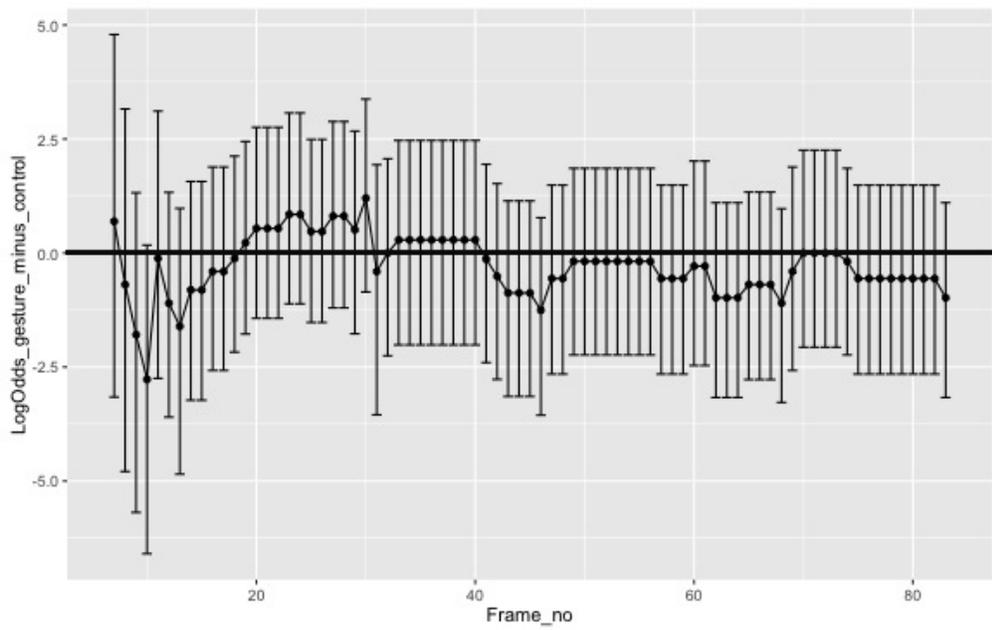


*Gesture condition*

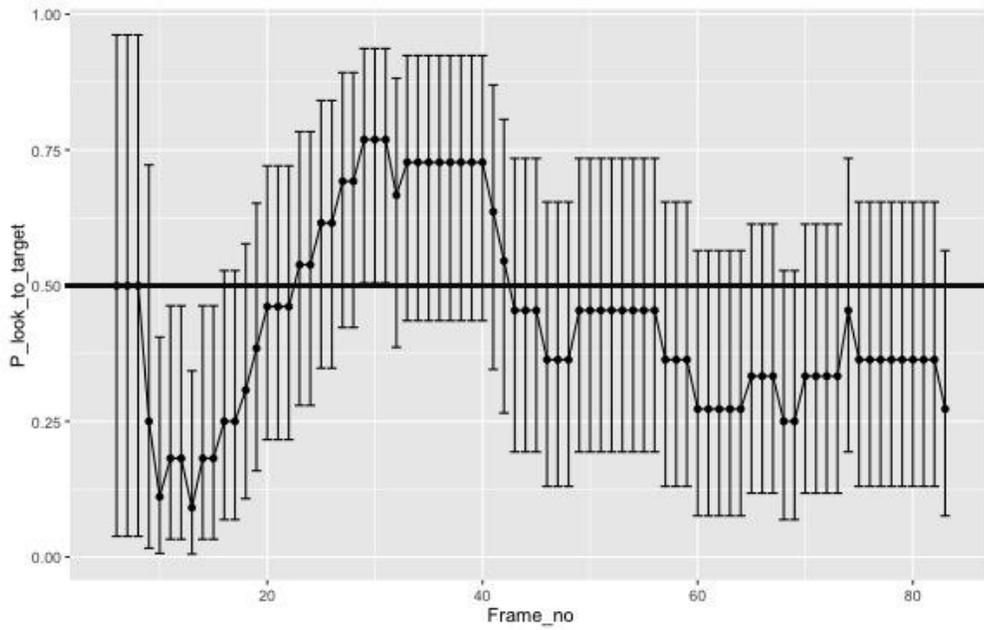


*Control condition*

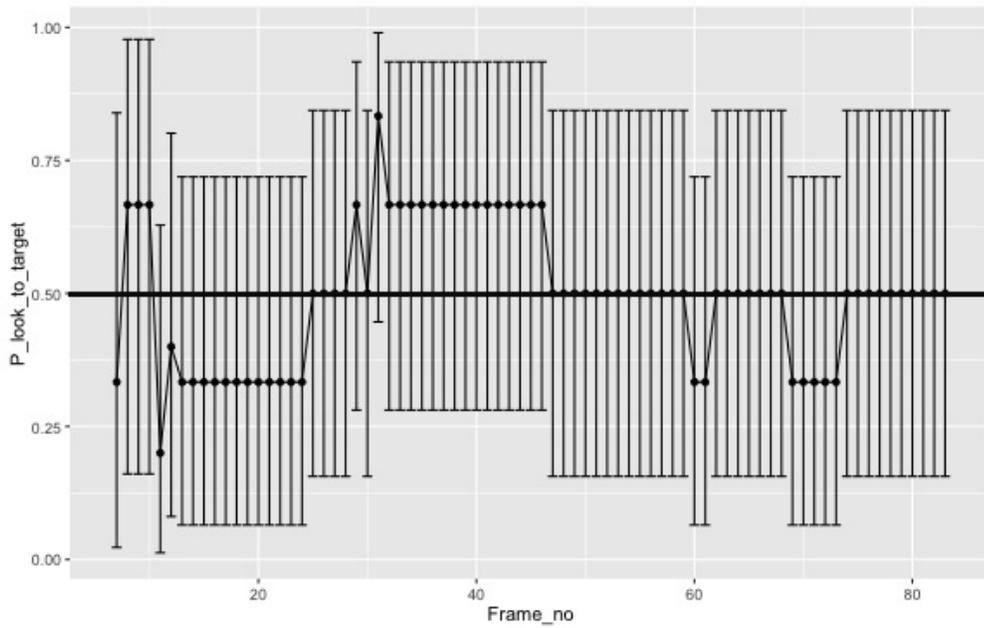
**Down**



*Gesture condition vs. control condition*

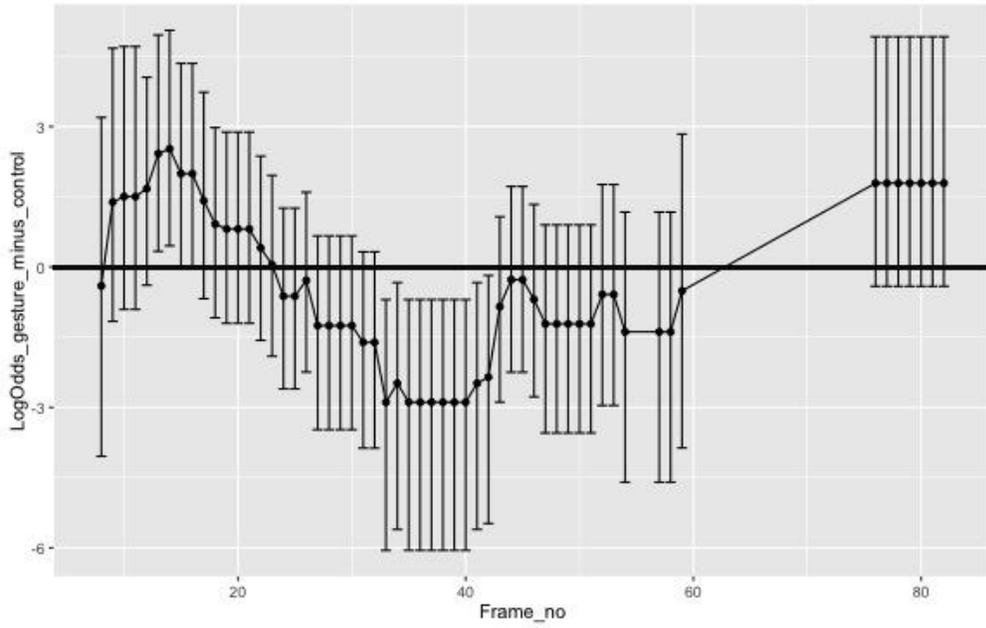


*Gesture condition*

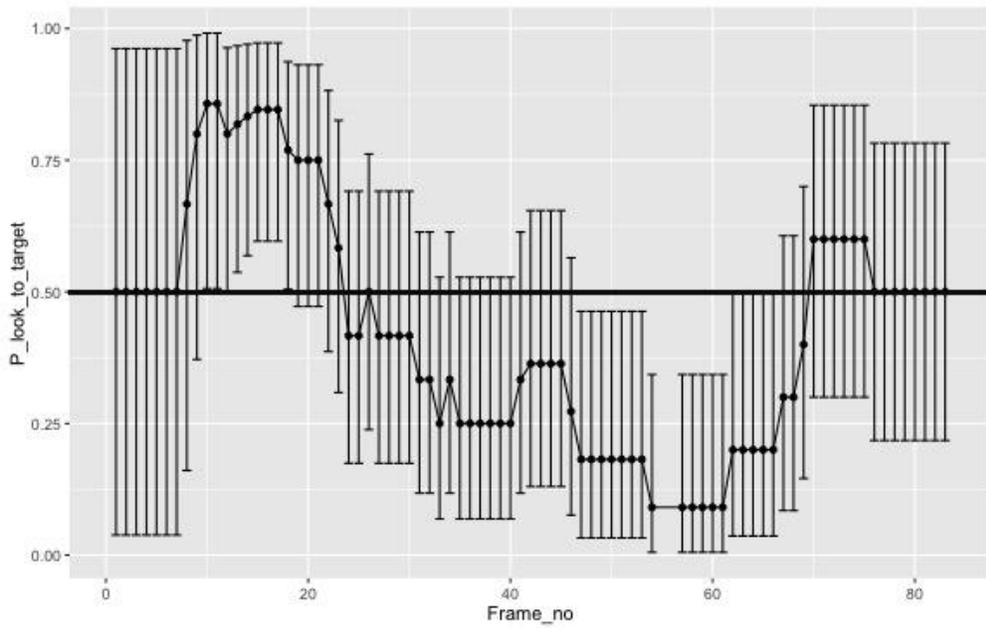


*Control condition*

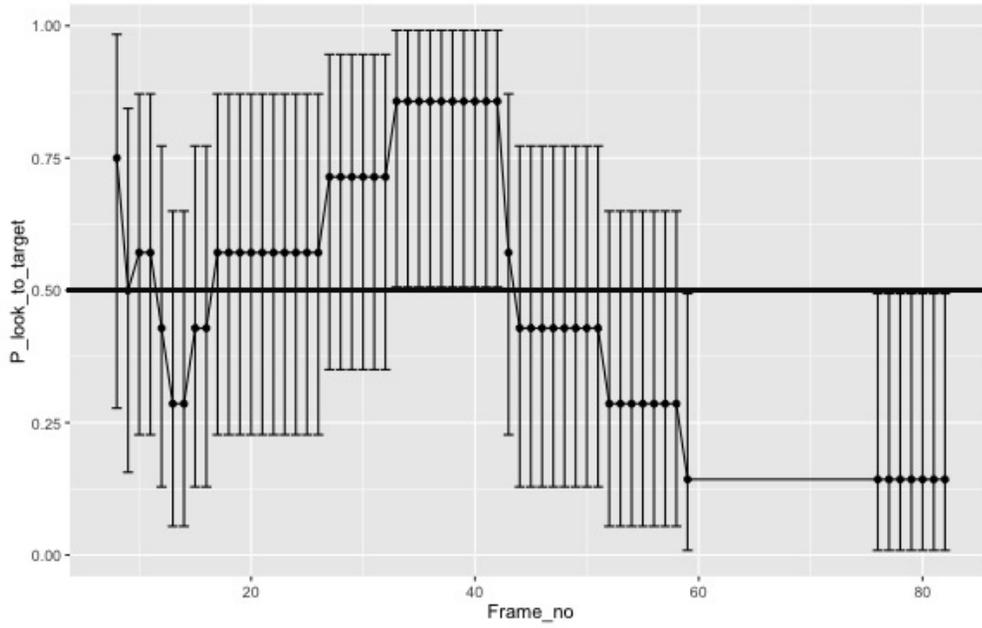
**Up**



*Gesture condition vs. control condition*

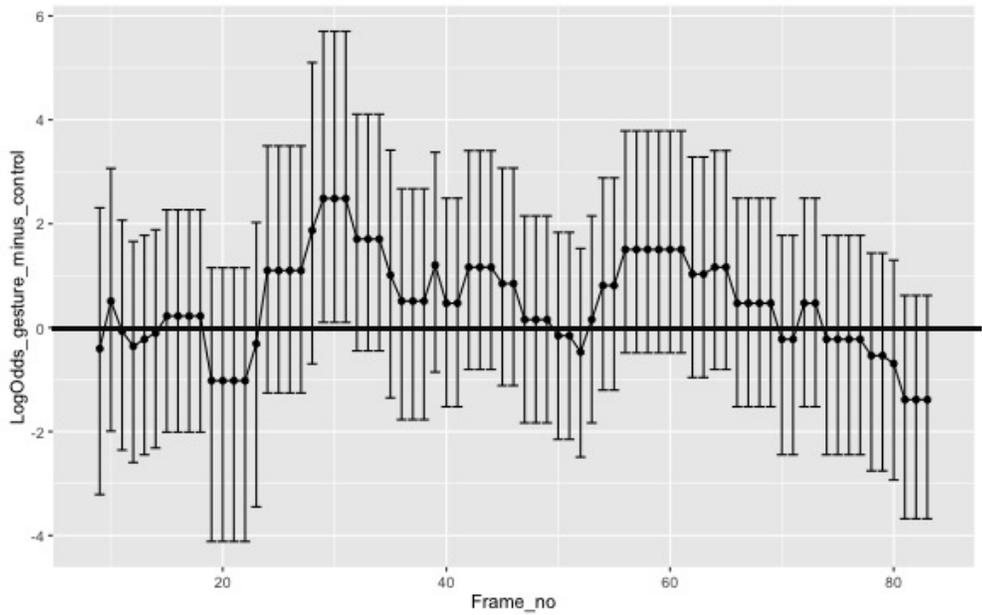


*Gesture condition*

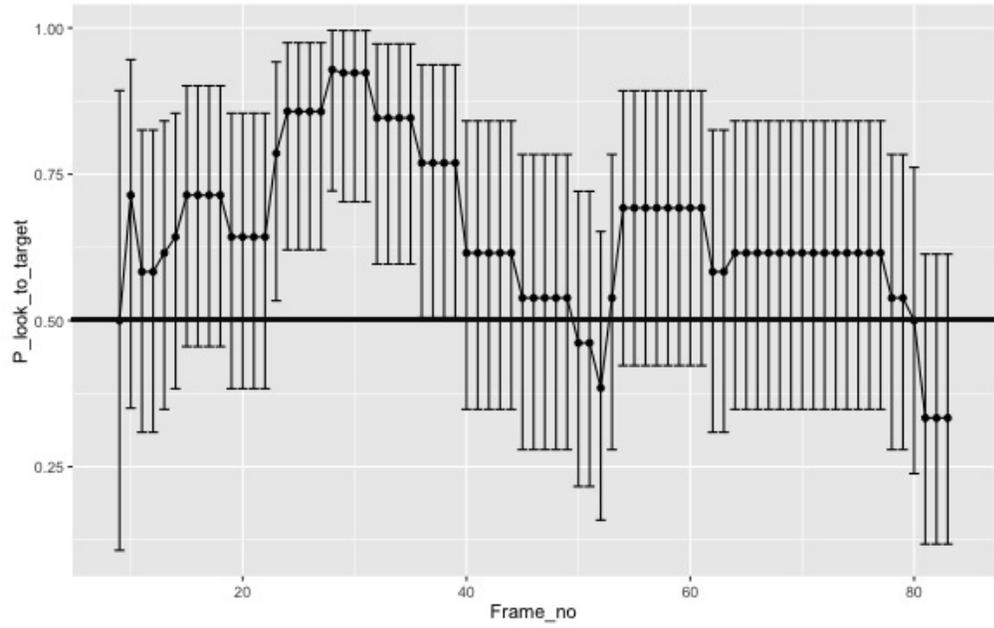


*Control condition*

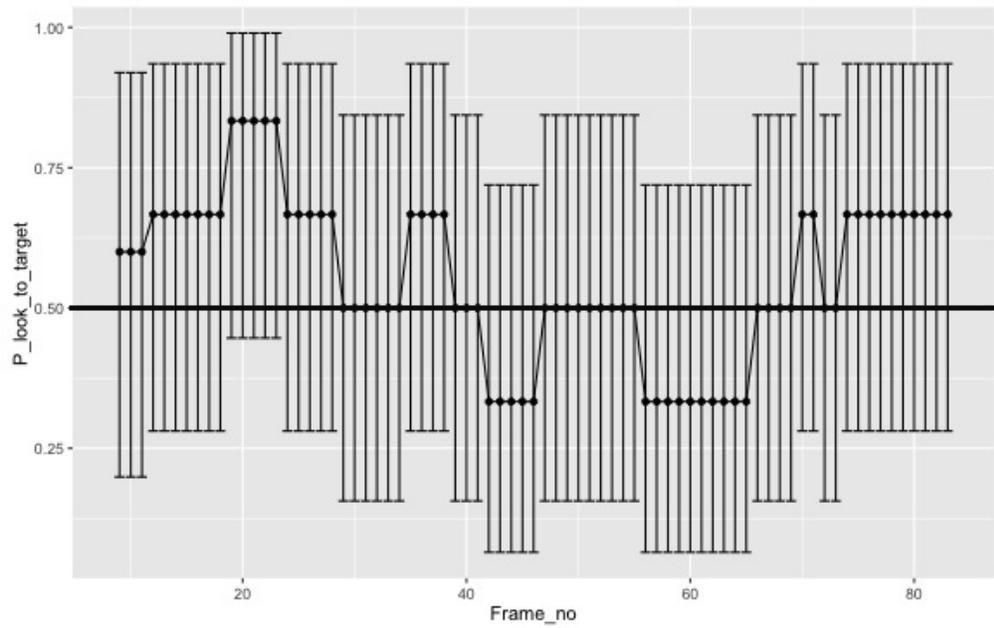
**Withdraw**



*Gesture condition vs. control condition*



*Gesture condition*



*Control condition*