Cognition 197 (2020) 104081

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Article

# Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking

Jessie S. Nixon

*Quantitative Linguistics Group, Eberhard Karls University of Tübingen, Germany*

ARTICLE INFO

*Keywords:*
Discriminative learning
Error-driven learning
Learning theory
Kamin blocking effect
Speech acquisition
Prediction error

ABSTRACT

Despite burgeoning evidence that listeners are highly sensitive to statistical distributions of speech cues, the mechanism underlying learning may not be purely statistical tracking. Decades of research in animal learning suggest that learning results from prediction and prediction error. Two artificial language learning experiments test two predictions that distinguish error-driven from purely statistical models; namely, cue competition – specifically, Kamin's (1968) 'blocking' effect (Experiment 1) – and the predictive structure of learning events (Experiment 2).

In Experiment 1, prior knowledge of an informative cue blocked learning of a second cue. This finding may help explain second language learners' difficulty in acquiring native-level perception of non-native speech cues. In Experiment 2, learning was better with a discriminative (cue–outcome) order compared to a non-discriminative (outcome–cue) order. Experiment 2 suggests that learning speech cues, including reversing effects of blocking, depends on (un)learning from prediction error and depends on the temporal order of auditory cues versus semantic outcomes.

Together, these results show that (a) existing knowledge of acoustic cues can block later learning of new cues, and (b) speech sound acquisition depends on the predictive structure of learning events. When feedback from prediction error is available, this drives learners to ignore salient non-discriminative cues and effectively learn to use target cue dimensions. These findings may have considerable implications for the field of speech acquisition.

## 1. Introduction

Listeners are able to discriminate remarkably fine-grained acoustic differences, when those differences discriminate meaning in listeners' (native) language(s). This is in stark contrast to differences in acoustic dimensions that are not discriminative, where perceptual sensitivity tends to be much lower. For example, as English is a non-tonal language, English native speakers tend to have lower sensitivity to the pitch dimension and therefore often have difficulty learning to use it as a lexical cue in tonal languages. Furthermore, there is enormous variability across languages in the types of acoustic dimensions, as well as the particular values of dimensions that discriminate meaning. Given this variability between languages, how do listeners learn which specific cues and dimensions are relevant for their own languages?

The present paper[1] begins with an extensive review of the literature, covering speech perception in infants and adults and two independent lines of research on learning: (a) human speech acquisition research, in particular distributional learning and (b) discriminative, error-driven learning, which has its origins in animal learning research, but has been increasingly being applied to human learning. Two key predictions are presented that distinguish between purely statistical and error-driven discriminative models – namely, cue competition (specifically, blocking) and unlearning. An introduction to the Rescorla–Wagner learning equations is presented, followed by two artificial language learning experiments with human participants, which test the above predictions. To preempt the results, the two experiments provide evidence that learning involves both cue competition and unlearning. While this finding does not rule out the possibility that other mechanisms, such as statistical learning, could occur in parallel, the results are difficult to explain with a purely statistical account.

### 1.1. Perceptual shaping as a result of linguistic experience

In the first months of life, infants are able to discriminate almost all of the speech sounds of the world's languages tested so far (e.g. Werker & Tees, 1984). However, as experience with the native language(s)

*E-mail address:* jessie.nixon@uni-tuebingen.de.

[1] The reference in the title to John Steinbeck's *Of Mice and Men*, which comes from Robert Burns' poem *To a Mouse*, is here intended simply to draw attention to the similarity in the learning mechanisms between people and other animals. Although the sentiment expressed in Burns' poem may also have a place here.

https://doi.org/10.1016/j.cognition.2019.104081
Received 19 March 2019; Received in revised form 18 September 2019; Accepted 18 September 2019

increases, the ability to discriminate non-native sounds decreases (Best, McRoberts, & Goodell, 2001). Adults often show poor discrimination performance with non-native sounds. Furthermore, in non-native speech perception, discrimination performance is affected by the *relationship* between native and non-native cues (e.g. Best, McRoberts, & Sithole, 1988; Flege, 1987, 1995). It appears that honing of the perceptual system to optimise native speech perception leads to poorer perception of acoustic differences that are not important for discriminating messages in the native language.

The general pattern of development of speech sound discrimination can be summarised thus: infants start out with acute sensitivity to small acoustic differences in almost every tested acoustic dimension used in human languages; adults often have sharpened sensitivity to acoustic differences that are meaningful in their native language, compared to infants; but adults also have reduced sensitivity in certain cases where differences are not meaningful. Specifically, sensitivity is reduced for acoustic dimensions that vary in speech but do not discriminate between lexical items (such as pitch in non-tonal languages). Sensitivity is also reduced within dimensions used for lexical discrimination when the particular range of cue values does not usually discriminate lexical items (i.e. reduced sensitivity to within-category differences). However, sensitivity is not reduced for sounds that do not occur at all in the native language (Best et al., 1988).

### 1.2. Distributional learning models

One highly influential approach to explaining the pattern of effects described above is the statistical (or distributional) learning approach. Speech production data show that acoustic dimensions exhibit statistical regularities that depend on whether the dimension is discriminative in a given language. For example, voice-onset time (VOT) forms two clusters (approximately Gaussian distributions) in languages such as English, that use VOT as a cue to lexical contrasts (e.g. Magloire & Green, 1999; Sundberg & Lacerda, 1999): a cluster of relatively short VOTs from multiple instances of *voiced* sounds in words such as 'bear', and another cluster of relatively long VOTs from *voiceless* sounds as in words such as 'pear'. This is in contrast to languages such as Pite Saami or New Zealand Maori, which do not use VOT as a lexical cue (Maclagan, Watson, Harlow, King, & Keegan, 2009; Wilbur, 2015). Statistical learning models propose that listeners track the statistical distribution of cues. In the above example, the number of VOT clusters would be used to determine the number of voicing categories: two (voiced and voiceless) in English and one in Maori or Saami.

Statistical learning models were proposed toward the end of the twentieth century in response to the dominant view at that time that speech was too complex to be learnable by general learning mechanisms and must therefore require special innate neural functions, such as "feature detectors" (see Eimas, 1985, for a review). Guenther and Gjaja (1996) presented a neural map model that was based on the statistical distribution of speech sounds. Non-uniformities in the acoustic distribution led to non-uniformities in neuron firing preferences in the auditory system. Further support for models based on statistical information in the input came from laboratory studies that showed that participants who are initially able to discriminate speech sounds *unlearn*, or learn to ignore, differences after exposure to a unimodal (one-cluster) distribution. For example, adult native speakers (Maye & Gerken, 2000) and 6–8-month-old infant learners (Maye, Werker, & Gerken, 2002) of English were exposed to either a bimodal (two-cluster) or unimodal distribution of a continuum from voiced [d] to voiceless unaspirated [t].[2] English-speaking adults and infants are able to discriminate these two sounds (Pegg & Werker, 1997). After exposure, compared to the bimodal groups, the unimodal groups in both studies were less likely to hear the endpoints of the continuum as different sounds. These studies suggested that, within the brief few minutes of an experiment, infants and adults can learn to ignore differences that they were originally able to discriminate. Furthermore, not only the number of clusters, but even specific properties of the distributions can affect the way people perceive speech sounds. For example, the statistical variance, i.e. the amount of acoustic variability within each category, affects how uncertain people are about what they are hearing (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Nixon & Best, 2018; Nixon, van Rij, Mok, Baayen, & Chen, 2016).

In recent years, a number of objections have been raised to the statistical learning hypothesis (see Cristià, McGuire, Seidl, & Francis, 2011; Werker, Yeung, & Yoshida, 2012, for reviews). For example, Werker et al. (2012) note that although effects have been found in very young, 6–8-month-old infants, older infants of 10–11-months do not reliably show distributional learning effects or may require substantially longer exposure (Yoshida, Pons, Maye, & Werker, 2010). There have been reported failures to find distributional learning effects (Terry, Ong, & Escudero, 2015; Wanrooij, de Vos, & Boersma, 2015) and the idea that learning depends on the number of peaks in a distribution has been challenged by the finding that, when dispersion was controlled, performance was equivalent between unimodal and bimodal distributions (Wanrooij, Boersma, & Benders, 2015). Computational models suggest that statistical learning alone may not be sufficient for learning the sounds of a language (Feldman, Griffiths, Goldwater, & Morgan, 2013; McMurray, Aslin, & Toscano, 2009). Furthermore, listeners' phonetic knowledge does not always correspond to the statistical distribution of acoustic information in speech.

### 1.3. Learning theory

Independently of the speech acquisition literature, the investigation of learning processes has a long history in animal learning research. In his seminal review, Rescorla (1988) argues that learning (often referred to as 'Pavlovian conditioning' in the animal learning literature) is the basis of how organisms come to represent the world and that what is essential is the *information* a stimulus (or cue) provides about the outcomes of events. Rescorla worried about a common misconception of Pavlovian conditioning as simply the forming of an association between two previously unassociated events – somewhat like a reflex response – resulting from their repeated co-occurrence or contiguity. The extent of this misconception is indicated in the title of his article, 'Pavlovian conditioning: It's not what you think it is' (Rescorla, 1988).

Learning theory and its formal implementations, most notably the Rescorla–Wagner model (Rescorla & Wagner, 1972), were developed to incorporate insights from decades of research in animal learning. Interestingly, researchers in Pavlovian conditioning also initially considered the idea that the strength of conditioning may depend on the statistical probability of cue–outcome co-occurrence (Rescorla, 1968). However, they soon found that co-occurrence statistics were not able to explain the pattern of results (Kamin, 1968, 1969; Rescorla, 1988; Rescorla & Wagner, 1972). In these models, rather than being probabalistic, learning is instead seen as error-driven. Learning results from all cues (referred to as the 'CS, conditioned stimulus,' in the animal learning literature) that are present in a given learning event competing to predict the relevant outcome ('US, unconditioned stimulus').

#### 1.3.1. Blocking and cue competition

One of the key findings that drove this new conceptualisation of learning and the development of the Rescorla-Wagner model was Kamin's (1968, 1969) demonstration that, at least in animal learning, cue learning can be diminished – 'blocked' – if a previously learned cue is sufficient for predicting a given outcome. Rats were trained with two cues, light + tone, co-inciding with electric shocks, then tested on a single cue (e.g. light). Prior to this two-cue training, half the rats

[2] This is not the canonical English aspirated /t/ as in 'top', but an unaspirated sound, created by removing the 's' portion from the beginning of the syllable, e.g. [ta] from 'sta'.

received pre-training with the cue that was *not* tested (e.g. tone). Rats who were not pre-trained showed a fear response (i.e. they 'conditioned') to the individual cue, e.g. the light. However, the rats who had been pre-trained with one cue did *not* condition to (i.e. did not learn) the second cue. Because the tone provided sufficient information for predicting shocks, the light had no additional predictive value and was therefore not learned during the two-cue training. This result shows that Pavlovian conditioning is not a simple case of statistical learning based on conditional probabilities, but instead a process in which all available cues compete in the process of predicting the relevant outcome. Since its publication, the Rescorla–Wagner model has become extremely influential in the field of animal learning and has helped explain a wide variety of observations (see e.g. Miller, Barnet, & Grahame, 1995; Siegel & Allan, 1996, for reviews).

#### 1.3.2. Error-driven human learning

Soon after it was published, the Rescorla–Wagner model began to also be effectively applied to human learning in areas as far-ranging as paired-associate word learning, category learning, correlational relationship judgments, transitive inference reasoning, social psychology, visual perception and physiological regulation (reviewed in Siegel & Allan, 1996). Feedback from prediction error also plays a vital role in the development of motor control (Cheng & Sabes, 2007; Shadmehr, Smith, & Krakauer, 2010) and human contingency learning (Dickinson, Shanks, & Evenden, 1984; Houwer & Beckers, 2002), as well as time perception (Ramscar, Matlock, & Dye, 2010) and pitch perception in music (Ramscar, Suh, & Dye, 2011). There is also neuroscientific evidence for the role of error in learning (see Schultz, 1998, for a review). Event-related potentials in a human contingency judgement task indicate that learning is error-driven, resulting from cue competition, rather than simple statistical tracking based on conditional probabilities (Kopp & Wolff, 2000).

Recently, a number of studies have specifically addressed the question of whether the predictions of error-driven learning models also apply in language (Apfelbaum & McMurray, 2017; Arnon & Ramscar, 2012; Baayen, Shaoul, Willits, & Ramscar, 2016; Chung, 2003; Colunga, Smith, & Gasser, 2009; Ellis & Sagarra, 2010; Olejarczuk, Kapatsinski, & Baayen, 2018; Ramscar, Dye, Gustafson, & Klein, 2013; Ramscar, Dye, & Klein, 2013; Ramscar, Dye, & McCauley, 2013; Ramscar, Dye, Popick, & O'Donnell-McCarthy, 2011; Ramscar & Yarlett, 2007; Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010; see Kapatsinski, 2018 for an excellent and thorough recent review). For example, Chung (2003) showed that for native English learners of Mandarin, learning of Chinese character meaning was faster and more accurate when the character preceded the translation during training, allowing participants to make predictions about the translation and learn from prediction error, compared to simultaneous presentation. Learning of pronunciation was also better with delayed feedback, compared to simultaneous presentation. The improvement with delayed compared to simultaneous presentation occurred both in the immediate post-training test and in follow-up tests two-weeks later. Arnon and Ramscar (2012) found effects of blocking during learning of determiner–noun pairings from prior learning of isolated nouns, compared to the reverse order in which whole sentences were learned first. Ellis and Sagarra (2010) found effects of blocking in English speakers learning Latin morphology.

Furthermore, while statistical learning models propose that learning is based directly on the actual, veridical distributions in the input, error-driven learning predicts instead that the greatest amount of learning will occur when there is surprise or prediction error. Therefore, in the distributional learning paradigm, the cues that occur more often (such as those near the mode of a Gaussian distribution) will each individually have less influence on learning than the rarer individual cues at the tails. Olejarczuk et al. (2018) recently showed that distributional learning of phonetic categories is driven more by the tails of the distribution, as expected with error-driven learning, rather than the veridical distribution predicted by distributional learning models.

### 1.4. Discriminitive learning and unlearning

While error-driven learning models have been highly successful in explaining many observed phenomena, some aspects of the models have tended to be missed in the literature. Perhaps most crucially, in focusing on the associations between events, theories have historically underplayed the importance of *unlearning* (see Ramscar, Dye, & McCauley, 2013, for a full discussion). Note that Rescorla and Wagner (1972) explicitly discuss the importance of unlearning, or decrements of associative strength, for uninformative cues. However, this aspect of the model often seems to have been overlooked. It is possible that this is due to the context and paradigms in which the model was initially formulated. For example, in Kamin's blocking experiments, which represent one of the most important findings that drove the development of the theory, an outcome (shock) occurred on every training trial. What happened to cue–outcome connections when the outcome was not present (i.e. negative evidence) was not explored. Neither did the issue of unlearning from negative evidence appear in Rescorla's (1988) review, although cue–outcome probability is discussed. Some authors (e.g. Xu & Tenenbaum, 2007) have specifically argued that learning does not require negative evidence (i.e. learning from absence of expected outcomes) and focus on modelling learning from positive evidence alone. However, recent evidence suggests that the unlearning that occurs with negative evidence is a critical component of error-driven learning, and may play an even greater role than positive evidence (Ramscar, Dye, et al., 2011; Ramscar, Dye, Gustafson, et al., 2013; Ramscar, Dye, & Klein, 2013; Ramscar, Dye, & McCauley, 2013; Ramscar & Yarlett, 2007; Ramscar, Yarlett, et al., 2010).

A particular variant of error-driven learning, *discriminative learning* (e.g. Ramscar, Dye, et al., 2011; Ramscar, Dye, Gustafson, et al., 2013; Ramscar, Dye, & Klein, 2013; Ramscar, Dye, & McCauley, 2013; Ramscar & Yarlett, 2007; Ramscar, Yarlett, et al., 2010) has been developed to address these issues and has been successfully applied to multiple areas of language processing, including semantic category acquisition, gender and plural acquisition, speech comprehension, morphological processing and lexical decision (Arnold, Tomaschek, Sering, Lopez, & Baayen, 2017; Arnon & Ramscar, 2012; Baayen et al., 2016; Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011; Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017; Ramscar, Dye, & McCauley, 2013; Ramscar & Yarlett, 2007; Ramscar, Yarlett, et al., 2010; Shafaei-Bajestan & Baayen, 2018).

The basic premiss of discriminative learning is that learning is a process of uncertainty reduction, rather than mere knowledge acquisition. In any learning event, any discriminable aspect of the environment can compete to predict relevant outcomes. Like other associative learning models, connection weights of cues that are (repeatedly) informative for predicting particular outcomes are strengthened. However, a critical aspect of discriminative learning theory is the role played by cue unlearning – the weakening or downweighting of uninformative cues – in forming a learner's knowledge of the environment. In any given learning event, weights are weakened between all present cues and any outcomes that are *not* present in the current input, if the outcome has been encountered before.

### 1.5. The Rescorla–Wagner learning equations

The predictions in the present study are derived from the Rescorla–Wagner equations, implemented in the `ndl` package (Arppe et al., 2015) in R (R Core Team, 2017). The Rescorla–Wagner equations estimate the connection strength, or *weights* $W$, between the input *cues* $C$ ($C \in c_k$, $k = 1, 2, ..., K$), and a set of *outcomes* $O$ ($O \in o_n$, $n = 1, 2, ..., N$). The network grows incrementally with each training trial. At the end of training with $k$ cues and $n$ outcomes, the network consists of a $k \times n$ matrix of connection weights. On each training trial, weights are adjusted between all and only cues present on that trial and all outcomes present or encountered previously. The

adjustment to the connection weight between a cue $c_i$ and outcome $o_j$ on a given trial, or learning event, $t$, is given by the Rescorla–Wagner equations:

$$w_{ij}^{(t)} = w_{ij}^{(t-1)} + \Delta w_{ij}^{t}.$$

The connection strength at the end of learning event $t$ is equal to the connection strength at the end of the previous learning event, $t-1$, plus any change during the current learning event.

The change in weights during the current learning event, $\Delta w_{ij}^{t}$, is given by the Rescorla–Wagner equations[3] :

$$\Delta w_{ij}^{t} = \begin{cases} \text{(a) } 0 & \text{if Absent}(c_i, t), \\ \text{(b) } \alpha_i\beta_j\left(\lambda - \sum_{[\text{Present}(c_k,t)]} w_{kj}\right) & \text{if Present}(c_i, t) \text{ and Present}(o_i, t), \\ \text{(c) } \alpha_i\beta_j\left(0 - \sum_{[\text{Present}(c_k,t)]} w_{kj}\right) & \text{if Present}(c_i, t) \text{ and Absent}(o_i, t), \\ \text{(d) } 0 & \text{otherwise} \end{cases} \quad (1)$$

in which $\lambda$ is the maximum learnability of the outcome; and $\alpha_i$ and $\beta_j$ refer to cue and outcome salience, respectively.

Put simply, the above equation says that (a) for any cue not present in a given learning event, no adjustment is made; (b) if a cue is present and an outcome is also present, the cue–outcome weight increases; (c) if a cue is present and an outcome is not present, cue–outcome weight decreases; (d) for any cues or outcomes that have not yet been encountered, no adjustment is made. The amount of adjustment made (in b and c) depends on the history of learning: the size of increase or decrease in strength is calculated based on the sum of connection strengths from the previous learning events of all present cues (subtracted either from $\lambda$ or from 0 and multiplied by the learning rate).

There are a couple of aspects of the model that are worth highlighting. Note that the model explicitly incorporates the notion that not only are weights increased when cues and outcomes co-occur, but for any cues present in a given learning event, weights decrease to outcomes that are not present, if those outcomes have been encountered previously. This is the formalisation of unlearning. Secondly, for all outcomes, the degree to which learning occurs depends on the history of learning of all cues present. When the total connection strength between all cues present and a given outcome is large (specifically, as it approaches lambda), then learning on positive trials becomes small. This is because the error remaining in the model is small. Conversely, if the total connection strength of all cues present is large and the outcome does not occur (negative evidence), there is a large amount of error and consequently of learning.

### 1.6. The present study

The present study investigates learning of non-native acoustic speech cues. The study concerns two related aspects of speech acquisition. Firstly, it relates to how speech sounds (cues) come to be used to effectively predict outcomes, such as word or morphological meanings. Secondly, it relates to how listeners learn which speech sound cues are important in their language and which are not – or gradient levels of importance – which is proposed to result from the same process.

Two main issues that differentiate discriminative learning models from statistical learning and positive-evidence-only associative models are tested. Key predictions of the three models are shown in Table 1. The first and second columns refer to distributions based on raw counts and cue competition during learning, respectively. In statistical learning models, a distribution is defined based on raw counts of all stimuli combined. (For example, frequency, unimodal or bimodal distributions of cues or co-occurrence probability.) Statistical learning models predict that learning will mimic the statistical structure of the input.

**Table 1**
Summary of the predictions of statistical learning, positive-evidence-only models and discriminative, error-driven learning models. Cue competition and blocking are investigated in Experiment 1; unlearning from negative evidence is investigated in Experiment 2.

| | Raw counts | Cue competition incl. blocking | Unlearning (negative evidence) |
|---|---|---|---|
| Statistical | Yes | No | No |
| Positive-evidence | No | Yes | No |
| Discriminative | No | Yes | Yes |

Associative models (both positive-evidence and discriminative), on the other hand, see learning as an iterative process, and therefore increment learning in a trial-by-trial manner. Outcomes are seen as having a limited maximum 'associative strength'. When the maximum associative strength is reached – that is, when an already-learned cue or cues perfectly predict an outcome – no more learning will occur. Therefore, associative models predict an effect of the history of learning on current learning (while statistical learning models do not). Due to the limited associative strength, cues compete for relevance in predicting an outcome, so if certain cues have already obtained a strong association with an outcome (i.e. already predict the outcome), any cues encountered later will be 'blocked' from being learned (see also Fig. 3 for an illustration of cue competition in the blocking effect). If blocking occurs during learning of speech sounds, this would be evidence against a purely statistical account of speech sound acquisition. In contrast, because statistical models are based on counts of data, learning can continue indefinitely and linearly. Cues do not compete for associative strength and, therefore, no effect of blocking is expected.

Experiment 1 tests whether speech sound acquisition involves cue competition in the form of blocking. Specifically, learning of non-native speech cues is expected to be better in the control condition, when there is no history of learning of the critical cues, compared to the blocking condition, when learning of the first cue during pre-training is expected to block learning of the second cue in the following training phase.

The third column of Table 1 relates to unlearning from negative evidence. Unlearning is the downweighting of connection strength between any present cues and any absent outcomes (specified in the third row of Eq. (1)). This is a key aspect that differentiates positive-evidence only models from discriminative learning. As discussed above, some researchers claim that learning is based on positive evidence alone. Even when this claim is not made explicitly, the importance of unlearning has often been overlooked in the literature. The focus has often been on positive association, although unlearning is inherent in the Rescorla–Wagner model. Even the name 'associative' suggests an emphasis on co-occurrence.

It may also be worth pointing out here that the Rescorla–Wagner equations assume an asymmetry between cues and outcomes. Cues predict outcomes. This means that the temporal relation between cues and outcomes is important. This aspect sets discriminative learning models apart from Hebbian models (e.g. Hebb, 1949), for instance, in which two cells simply 'fire together'.

Experiment 2 tests whether acquisition of speech sounds involves unlearning. Learning of acoustic cues is expected to be better when the acoustic cues precede the visually presented semantic outcomes, compared to the reverse order. This is because some of the acoustic cues are informative for predicting the outcomes and some are not. When the cues occur before the outcomes (discriminative order), predictions can be made about which outcome is expected to occur based on the various acoustic cues. When uninformative cues fail to predict an outcome, this leads to prediction error and the connection weight between these cues and the outcome is weakened – that is, the cues are *unlearned* as

[3] A simplified description of the model is presented here. For a more formal version, see Appendix A.

predictors of that outcome. On the other hand, when informative cues accurately predict an outcome, connection weights are strengthened. Over time, this process of cue competition leads to stronger connection weights for informative cues, compared to uninformative cues (see also Fig. 8 for an illustration of unlearning and the effect of cue–outcome order).

However, when the visual outcomes precede the acoustic cues (non-discriminative order), there is no cue competition. The visual shapes are simple with no variation, so there are no cues that can compete with each other. Each shape occurs with two different acoustic stimuli with a certain probability. Because there is no cue competition, connection weights simply fluctuate, increasing when an outcome occurs and decreasing again when it does not occur. Therefore, participants are expected to learn the *probability* of encountering each of the acoustic stimuli. Performance on high-frequency stimuli will appear good, at least on the surface. Because of the high frequency of occurrence (a certain shape with a certain syllable), even if responses are based on the wrong (non-discriminative) aspects of the syllable, they will still be correct most of the time. However, in the low-frequency stimuli, if responses are based on the non-discriminative aspects of the syllable, they will be wrong.

Therefore, no difference between conditions is expected for the high-frequency stimuli, but learning of low-frequency stimuli is expected to be better in the discriminative order, compared to the non-discriminative order. Note that the stimuli are identical between conditions; only the order of acoustic cues versus visual semantic outcomes is manipulated.

## 2. Experiment 1: Blocking

Statistical learning models propose that learning is based on the statistical distribution of all cues collected together (see Fig. 1). In particular, a bimodal distribution of acoustic cues should lead to discrimination of two sounds due to the forming of two clusters (Maye et al., 2002; Maye & Gerken, 2000; Maye, Weiss, & Aslin, 2008). This is based on the idea that listeners form “mental histograms” by keeping track of the frequency of occurrence of cue values (Maye & Gerken, 2000). Importantly, because statistical learning is based on frequency counts, the order in which cues are learned does not affect learning.

Error-driven learning models, on the other hand, propose that in each learning event – or trial – all cues in the input ‘compete’ in the process of predicting the outcome. This means that learning on each trial depends on the *history of learning* of the cues. Only when there is uncertainty about the outcome is there potential for learning about the present cues. When a strong predictive relationship has already been formed between a cue and outcome – that is, when the outcome is highly predictable – there is no longer room for further learning. In the Rescorla–Wagner model, learning has reached asymptote. Kamin described this same concept in terms of ‘surprise’. Changes in cue–outcome connection strength only occur when the outcome leads to surprise. When an outcome is predictable from already-learned cues, then when the outcome does occur, it is expected and there is therefore no surprise – that is, no *error* – to drive further learning, including learning of any other cues that might be present. To illustrate using the case of Kamin's rats described above, after the first training phase (light then shock), whenever the rats saw the light they *knew* the shock was coming. There was no surprise or uncertainty left to drive learning of the new cue, tone, in the second (light + tone then shock) phase.

Predictions are shown in Fig. 2 for discriminative learning. The discriminative learning model (R package `ndl`) is initially presented with a single cue (e.g. tone) or control cue (VOT) to an outcome (size), then later presented with two cues (nasality + tone) to the same outcome. The plot shows predictions for learning the second cue (nasality). The model predicts that learning will be poorer after blocking pre-training, compared to control pre-training.

In Experiment 1, participants saw two pictures on the computer screen and heard a spoken word. They were instructed to click on the picture corresponding to the word. The two pictures were the same except that one was the original size and one, the ‘diminutive’, was smaller. Participants had to learn the speech cue signalling the diminutive. Just like the model, human participants were presented with a single cue (e.g. tone) or a control cue (VOT) to an outcome (size) during the pre-training phase. Next, in the training phase, they were presented with two cues (nasality + tone) to the same outcome. In the test phase, participants were tested only on the cue that did not occur in the pre-training phase (e.g. nasality).

If listeners learn the speech cues by discriminative learning, we expect accuracy to be lower after blocking pre-training, compared to control pre-training (Fig. 2). If listeners learn from statistical distributions, we would expect no difference between conditions for the second cue, as the two categories should emerge from the two clusters of acoustic cues, which are the same between conditions (see Fig. 1).

### 2.1. Method

Both experiments used the ‘roofrunner’ online game (Rácz, Hay, & Pierrehumbert, 2017; see also Beckner, Pierrehumbert, & Hay, 2017; Schumacher, Pierrehumbert, & Lashell, 2014, for different game variants). A ‘flying creature’ communicates with an interlocutor in order to continue flying along the rooftops. Training phases had a daytime setting and test phases a night-time setting.

#### 2.1.1. Participants

Participants in Experiment 1 were 187 native speakers of English living in the US, recruited online via Amazon Mechanical Turk. They were paid $4 for participation. In both experiments, all participants

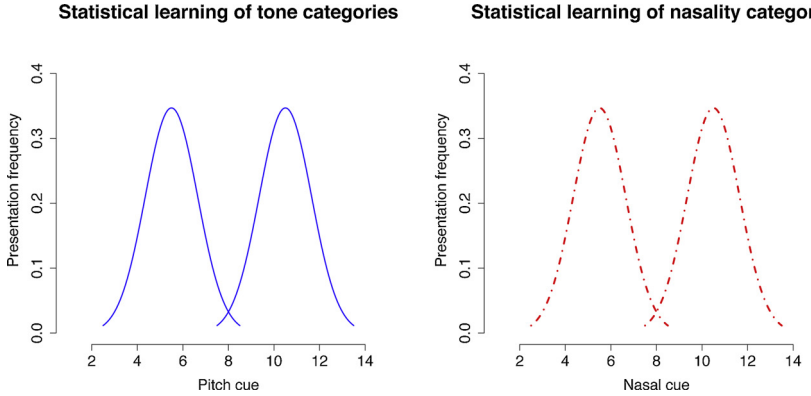


**Fig. 1.** In statistical learning models, learning is based on the frequency distribution of acoustic cue values, for example, pitch in tone categories (left panel) or nasality in oral versus nasal vowels (right panel).

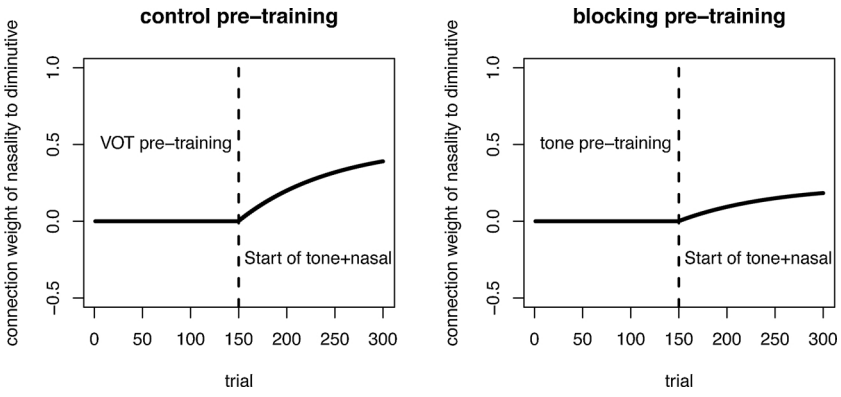


**Fig. 2.** Simulation of the Kamin blocking effect applied to acoustic cue learning, as in Experiment 1. The connection weight (*y*-axis) over trials (*x*-axis) is shown for the nasality cue to the diminutive outcome in the control (VOT) pre-training condition (left) and blocking pre-training condition (right). In the blocking condition, a single cue is first learned as a predictor to a particular outcome, in this case tone as a cue to diminutive size (pre-training phase). Later (Phase 2; onset indicated by the vertical dashed line), a second cue is presented simultaneously with the original cue: tone and nasality as cues to diminutive size. The control condition is the same except that a control cue (VOT) is presented in the pre-training phase, instead of a critical cue. In Phase 2, the connection weight of nasality increases in the control condition, but remains low in the blocking condition. In the blocking condition, learning of the second cue, nasality, is 'blocked' by the first cue, tone. The simulation assumes equal salience of tone and nasality cues.

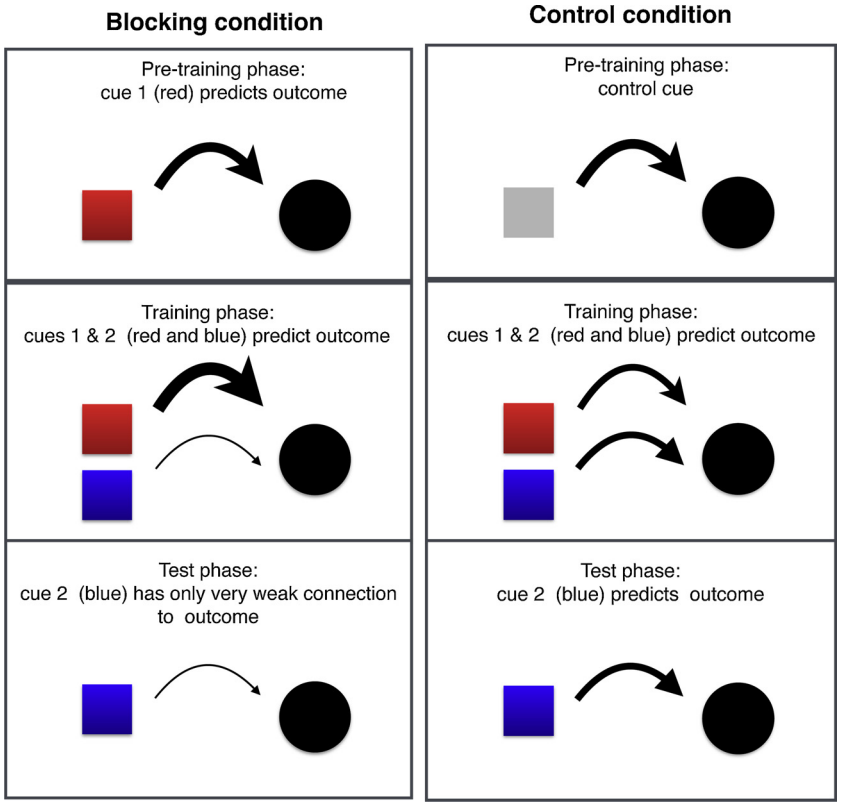


**Fig. 3.** Illustration of cue competition in the blocking effect. Arrow thickness indicates (very approximately) connection strength. **Blocking condition (left).** In the *Pre-training phase*, the first cue (red) predicts the outcome (circle) and connection weights develop (indicated by the arrow). In the *Training phase*, two cues (red and blue) are presented. Connection weights from both cues (red and blue) to the outcome increase. However, the *amount* of increase is small. Because the maximum connection strength to the outcome is limited, and because the red cue has already been learned, there is less connection strength left for learning in the training phase. The amount that is left is split between the two cues. Therefore, at the end of the training phase, the connection weight from the second cue (blue) is still weak. During the *Test phase*, if the blue cue is presented, it only weakly predicts the outcome. **Control condition (right)**. In the *Pre-training phase*, there is no learning of the red cue in the control condition. Only a control cue is learned. Therefore, in the *Training phase*, both cues (red and blue) are learned together at the same rate, reaching equal connection strength. In the *Test phase*, the blue cue is therefore stronger than in the blocking condition. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

gave consent in an online informed consent form. After the experiment, participants filled out a questionnaire about their demographic background, including questions about their native language and in which country they learned this language. Participants whose native language was not English acquired in the US were excluded from the analysis (eight participants). An additional 22 participants were also recruited but were excluded from analysis because they did not complete the experiment (20 participants) or due to technical problems that led to them completing too many trials (two participants). Each experiment lasted approximately 20–30 min.

### 2.1.2. Stimuli

In both experiments, auditory stimuli were monosyllabic words produced in isolation by a native speaker of Southern Min with phonetics training. Southern Min was selected because it is a tonal language with a nasality distinction. In Experiment 1, this meant that two acoustic dimensions that are non-native for English speakers could be independently manipulated with both cues occurring in the vowel to match the timing of the cue as closely as possible. Naturally produced speech tokens were used to avoid artefacts in synthesised speech. Naturally produced speech tokens are also more speech-like, more likely to be processed as language and therefore may be more learnable (Logan, Lively, & Pisoni, 1991), compared to artificially produced stimuli. Sixty-four different base (segmental) syllable types, two tones (mid-level, high-level) and two levels of nasality (nasal, oral) were used. Different segmental syllables were used between pre-training, training and test.

Mid-tone, oral words were used as *baseline* stimuli, because they are most similar to English speech. The more acoustically distant non-native speech sounds are from native language speech sounds, the more likely they are to be detected and perceived, while similar sounds tend to be 'assimilated' to native speech sounds (Best, 1995; Best et al., 2001; Flege, 1995; Flege, Takagi, & Mann, 1995). The mean fundamental frequency (f0) for the different cue types was: mid-tone oral: 198 Hz; high-tone oral: 242 Hz; mid-tone nasal: 198 Hz and high tone nasal: 239 Hz. Because mid-level tones are produced around the middle of the

**Table 2**
Experiment design of Experiment 1. The experiment has two conditions, which depend on the type of pre-training (*control*: with control cues, voice-onset time, VOT; or *blocking* with nasal or tone cues). Pre-training phase: in the blocking condition, participants hear a critical cue (tone: low versus high; or nasality: oral versus nasal). In the control condition, they hear a control cue, voice onset time. Training: in the following training phase all participants hear two cues (tone plus nasality: low, oral versus high, nasal). Finally, in the test phase, they hear the cue that was not pre-trained (nasality or tone).

| Condition | Pre-training | Training | Test cue |
|---|---|---|---|
| Blocking | Tone | Tone + nasal | Nasality |
| Control | VOT | Tone + nasal | Nasality |
| Blocking | Nasality | Tone + nasal | Tone |
| Control | VOT | Tone + nasal | Tone |

voice range, they are close to the f0 used in US English speech. For native US English speaking women mean f0 is around 200 Hz (Pépiot, 2014). The high-level tone, in contrast, is higher than typical US English speech and therefore more likely to be detectable as an additional cue, compared to the mid-level tone. Similarly, oral vowels are most similar to English speech, as English does not have a nasality distinction for vowels. Therefore, the low oral vowels are likely to be perceived as more typical English sounds and the nasal vowels and high-level tones are more likely to be perceived as additional cues. Accordingly, low oral words were set as the baseline and corresponded to the baseline (original size) images, while the additional cues, high pitch and nasality corresponded to the diminutive images.

The visually presented semantic stimuli were 128 images adapted from the game Glitch by Tiny Speck (http://www.glitchthegame.com/public-domain-game-art/). Pairs consisted of original and reduced-size ('diminutive') images (45% of the size of the original). Each image pair was presented twice, once for each corresponding word.

#### 2.1.3. Experiment design

Experiment 1 had three phases: pre-training, training and test (see Table 2). In pre-training, participants received either *blocking pre-training* or *control pre-training*. The following training phase was identical for all groups. The test cue was counterbalanced between participants. Therefore, there were four participant groups: 2 pre-training conditions (blocking, control) × 2 cue types (tone, nasality).

Participants receiving blocking pre-training were trained with one critical cue (e.g. tone). The baseline (mid-tone oral) corresponded to original sized images; additional cues (e.g. high tone) corresponded to diminutive images. Participants receiving control pre-training were trained with a control cue, VOT, that was not used in test. In the training phase, all groups received identical training: both cues simultaneously (i.e. high, nasal stimuli) signalled the diminutive. In the test, participants were tested on the cue that was not pre-trained (e.g. nasality).[4]

The prediction is that, if blocking occurs in speech learning, performance will be better in the control condition than in the blocking condition. In the blocking condition, learning of one cue during pre-training should block learning of the second cue during the training phase. So, when the second cue is presented during the test phase, performance will be worse after blocking pre-training than after control pre-training.

#### 2.1.4. Trial procedure

In both the pre-training and training phases, participants saw two images on the screen: one original size, one diminutive. An auditory word was presented 200 ms prior to the images. There was an inter-trial interval of 1000 ms. The order of trials was randomised individually for each participant. The location of the images on the screen (left or right) was also randomised throughout the experiment. In the first half of the pre-training and training phases, the two corresponding trials (original size and diminutive) occurred in sequence. The task was to click on the image corresponding to the word. Feedback was given on each trial by highlighting target pictures in green/red for correct/incorrect responses. Additionally, the character flew forward on correct trials, but was knocked down on incorrect trials. During the test, the trial procedure was the same except that corresponding trials were not presented in sequence and feedback was not given.

### 2.2. Analysis and results

A generalised linear mixed effects (glmer) model tested effects of pre-training on probability of selecting the target (using the `lme4` package in R; Bates, Mächler, Bolker & Walker, 2015). The model included three two-level factors: image size (original, diminutive), cue type (tone, nasal) and pre-training condition (blocking, control) and their interactions. Random intercepts for participants and items and random slopes for size by participant were included. Trial was tested, but did not improve model fit so was removed. Random slopes for condition by item were tested, but the models did not converge.[5]

Fig. 4 shows model estimates of the proportion of correct responses in the blocking versus control conditions for diminutive and original size images (left panel) and for tone and nasal stimuli in the control versus blocking conditions (right panel). Table 3 shows the model summary. Original size targets were selected significantly more often than diminutives. This is probably because they were baseline stimuli and would be the default if additional cues were not perceived. Size also interacted with cue type. For diminutive items, targets were selected more often for the tone cue than the nasal cue. There was a significant three-way interaction between size, condition and cue. For original size stimuli, there was no significant difference between conditions (nasal $p = 0.168$; tone $p = 0.085$).[6] Most importantly for the present study, for the nasal stimuli, as predicted, participants selected diminutive targets significantly more often after control pre-training than blocking pre-training. There was no effect of training type for tone stimuli ($p = 0.483$). The significant effect of blocking for the nasal cue supports the prediction that prior knowledge of an informative cue can block learning of a second cue – even when that cue would otherwise be informative for predicting the outcome, as the control condition shows. This replicates Kamin's (1968) finding of a similar effect on stimulus conditioning in rats.

One question the results raise is why the blocking effect was not significant for the tone stimuli. The overall accuracy for tone stimuli was significantly higher than nasal stimuli ($p < 0.0001$). This suggests that the absence of blocking for tone may be due to the nasal cue not being fully acquired during pre-training. In order to further investigate the validity of this explanation, accuracy during the pre-training phase

[4] An alternative control condition was initially considered; namely, no pre-training. However, this possibility was decided against, on the basis that if a blocking effect were found, it could be argued that this could be due to fatigue, boredom or lack of concentration due to having twice as many trials in the blocking condition. A control condition with control cues was selected instead, to control for the number of trials between conditions. This also arguably provides a stronger test of the blocking effect, since it might be expected that participants would do worse with control cues compared to no pre-training, due to the potentially confusing task of learning one set of control cues and then having to learn a different set of cues. However, it is also possible that the different cue used in the control pre-training could alert participants to the possibility that cues can vary, enhancing sensitivity to changes in cues. This possibility could be tested in future experiments with a no-pre-training control condition.

[5] Convergence failure occurred regardless of whether dummy or effects coding was used.

[6] Obtained by relevelling the model.

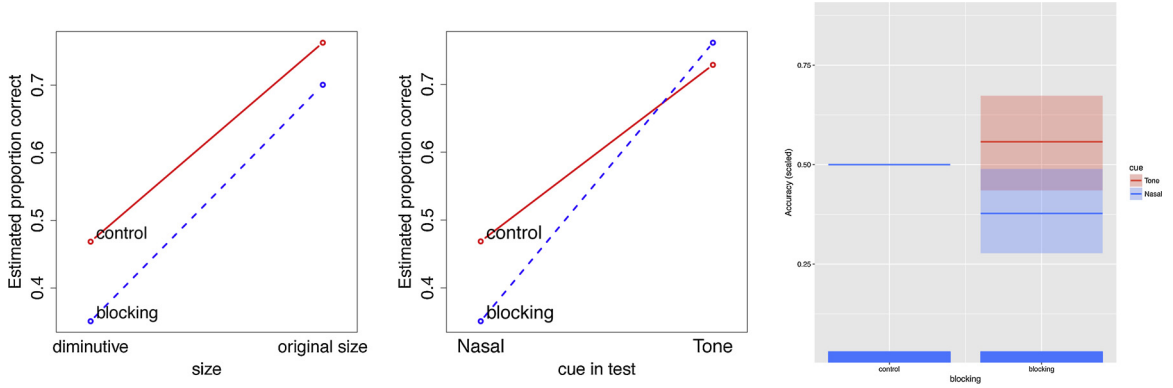


**Fig. 4.** Model estimates for the results of Experiment 1. Left: estimated proportion of correct responses for the blocking (blue, dashed line) versus control condition (red, solid line) for diminutive and original size pictures. Centre: estimated proportion correct for the nasal (left) and tone cues (right) in the blocking (blue, dashed line) versus control condition (red, solid line) for diminutive pictures. Right: Plot of the estimates of the contrast of accuracy in the blocking condition relative to the control condition. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 3**
Model summary for effect of blocking versus control pre-training on selecting the target picture during the test phase. Model is dummy coded.

| Fixed effects | Estimate | Std. error | $z$-value | $Pr(<\|z\|)$ |
|---|---|---|---|---|
| (Intercept) | −0.1304 | 0.1840 | −0.709 | 0.4836 |
| Cond = blocking | −0.5006 | 0.2329 | −2.150 | 0.0314 |
| Size = original | 1.3544 | 0.1329 | 10.188 | < 0.0001 |
| Cue = tone | 1.1197 | 0.2598 | 4.310 | < 0.0001 |
| Cond = blocking: size = original | 0.1545 | 0.1249 | 1.237 | 0.2094 |
| Cond = blocking: cue = tone | 0.7313 | 0.3421 | 2.137 | 0.0329 |
| Size = original: cue = tone | −0.7093 | 0.1615 | −4.393 | < 0.0001 |
| Cond = blocking: size = original:cue = tone | −0.7212 | 0.1820 | −3.962 | < 0.0001 |

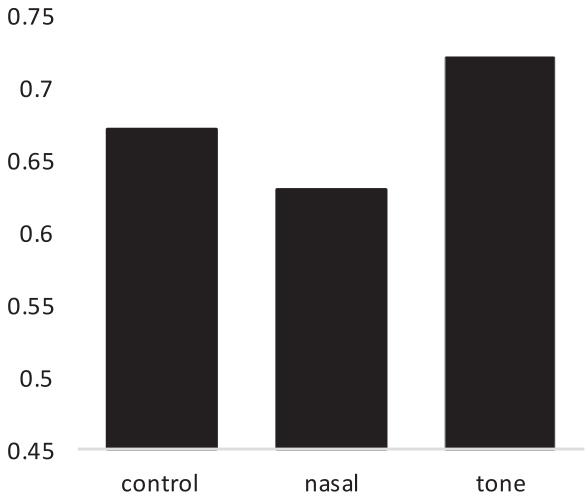


**Fig. 5.** Pre-training phase accuracy across conditions.

was inspected. The pre-training accuracy across conditions is shown in Fig. 5. Learning is comparable across the blocking and control conditions during the pre-training phase. A separate glmer model of pre-training accuracy was run. The model showed no difference in accuracy between the tone control and the nasal control groups ($t = 0.03$); however, accuracy was significantly higher for the tone stimuli than the nasal stimuli during the pre-training phase ($t = 2.68$). If the nasal cue did not do a good job of predicting the image outcome for these listeners by the end of pre-training, then there would still be sufficient uncertainty during the following training phase to drive learning of the new cue (Rescorla & Wagner, 1972).

Inspection of the effect of individual pre-training scores on performance in the test phase provides further evidence for a blocking effect. A separate glmer model was run on the data from the blocking condition. Accuracy during the pre-training phase was used as a numerical predictor of accuracy during the test phase. The model included cue type, size and pre-training score, and the interaction between size and pre-training score, as well as random intercepts for participants and items. Pre-training score was significant ($z = -2.546$, $p = 0.0109$) as was the interaction between pre-training score and size ($z = 14.318$, $p < 0.0001$). When the interaction with cue type was included, the model failed to converge; however, the effect of pre-training score seems to stem from the nasal cue. The model results are shown in Fig. 6. The results showed that the better the score during pre-training, the greater the blocking effect; that is, the better the pre-training score, the worse the performance during test for the diminutive items and the better the score for the baseline original size items.

### 2.3. Discussion

Experiment 1 tested whether acquisition of speech sounds involves cue competition, as predicted by error-driven learning models. Either a single critical cue (e.g. tone; blocking condition) or a control cue (control condition) was presented during the pre-training phase, then

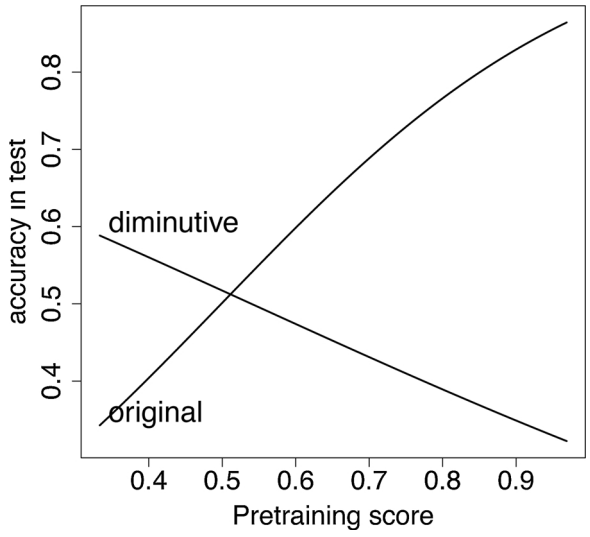


**Fig. 6.** Model plot of effect of pre-training score on accuracy during the test phase for diminutive and original size items. The better the pre-training score, the worse the performance for the diminutive items and vice versa for the original size items.

this critical cue was presented along with a second cue (e.g. nasal) in the next training phase. In the test only the second cue was presented. Accuracy was significantly higher after control pre-training than blocking pre-training for the nasal cue. When the first cue was learned in the pre-training phase, this 'blocked' learning of the second cue. These results replicate Kamin's (1968) blocking effect, originally demonstrated in rats, and demonstrate that cue competition also occurs during learning of speech cues.

In addition to the effects of the blocking condition, inspection of the individual scores in the pre-training phase on accuracy during the test phase provides further evidence for blocking. This negative correlation between pre-training and test accuracy is particularly striking considering that it goes in the opposite direction to what might be expected on the basis of individual differences in perceptual ability.

The present results are difficult to reconcile with a purely statistical clustering account of speech sound acquisition, because the statistical distributions were the same in both conditions. According to statistical learning, two categories of speech sounds should form in both conditions, due to the bimodal distribution (see Fig. 1). Note that these results do not rule out the possibility that statistical tracking also occurs, perhaps in parallel. However, statistical tracking alone cannot account for these effects. Importantly, under certain circumstances, namely when discriminative cue structure – and therefore, cue competition – is not available, error-driven learning predicts that learning will reflect the statistical properties of the input (as we will see in Experiment 2).

The finding of a blocking effect for speech cue learning has implications for first and second language acquisition and speech perception. As noted above, language acquisition changes over the course of development. In experimental paradigms in which infants unlearn speech sound differences that they can initially discriminate, 10–11-month-old infants require longer exposure than 6–8-month-old infants (Werker et al., 2012; Yoshida et al., 2010). Especially as adults, but even as infants, the experience of learning our native language affects our use of non-native languages. Our perception and production of non-native sounds depends on the relationship of the non-native sounds to our native language (e.g. Best et al., 1988; Flege, 1987, 1995). The *history of cue learning* may affect speech acquisition when already-learned native cues block learning of new, non-native cues.

It appears that tone was easier to learn than nasality. This difference in learnability of the tone and nasal cues may be due to prior language exposure. Although English uses nasal consonants (Turnbull, Seyfarth, Hume, & Jaeger, 2018), it does not use oral versus nasal vowels to discriminate word meaning, as in Southern Min. For English speakers, the most likely encounter with nasal vowels is probably from French, either from learning French as a second language or from cultural exposure, such as in films and other media. In French, nasal vowels co-occur with a shift in formant frequency; for example, the words 'francais' (e.g. 'Parlez-vous francais?'), 'blanche' (e.g. 'carte blanche'), the name 'Jean' (e.g. 'Jean-Luc Picard'). The vowel quality changes along with the nasality. This may lead to a strong expectation in English speakers that nasal vowels have a shift in the vowel quality relative to their oral counterparts, making it more difficult to detect a nasal vowel without a shift in the vowel formants. Another possibility is that pitch may be perceived as indicating diminutive by English listeners (Ohala, 1983; Ohala, Hinton, & Nichols, 1997; Tsur, 2006, see also Monaghan, Shillcock, Christiansen, & Kirby, 2014).

As mentioned above, discriminative learning predicts that when cue competition is not available, learning will appear more like statistical learning and more closely resemble the statistical input. Experiment 2 tests how the presence versus absence of cue competition affects learning of speech sounds. The effects of cue competition are tested by manipulating the temporal order of complex, discriminative acoustic stimuli, which contain predictive cue structure, and simple, non-discriminative visual outcome stimuli, which do not contain predictive cue structure.

The Kamin blocking effect is an iconic symbol of error-driven learning and cue competition. Kamin's finding was one of the key contributions to the development of error-driven learning theory and the Rescorla–Wagner model. Therefore, the design of Experiment 1 was closely modelled on the original blocking experiments. However, it is also important to note that Kamin's experiment was only designed to test increases in connection strength through co-occurrence of cues and outcomes. There were no trials where a cue that had been previously associated with the shock later occurred in the *absence* of the shock. So, it did not test weakening of connection strength when cues and outcomes do not co-occur. Consequently, Experiment 1 does not tease apart whether learning speech sounds occurs only when cue and outcome co-occur (positive evidence) or whether unlearning occurs for connections between present cues and absent outcomes (negative evidence), as predicted by the Rescorla–Wagner equations. This is investigated in Experiment 2.

## 3. Experiment 2: Cue–outcome order

Experiment 2 tests two related questions. Firstly, does learning of speech sounds occur simply due to 'association' of two previously unrelated events (such as an acoustic stimulus and a semantic stimulus)? Or is the relationship *predictive?* If it is predictive, this means an asymmetrical relationship, in which *cues*, which occur earlier in the temporal sequence of events, predict *outcomes*, which occur later. If it is simply an association, the temporal sequence of the stimuli should not matter. If it is predictive, the temporal order of events should affect learning whenever there is not a one-to-one mapping between cues and outcomes.

Secondly, Experiment 2 tests whether learning of speech cues is based only on positive co-occurrence between cues and outcomes or whether cues are also *unlearned* when cues are present and outcomes are not present. If learning is based only on positive co-occurrences, we should see learning of the statistical structure. That is, participants should learn the probability with which particular cues are associated with particular outcomes. If only positive co-occurrence is important, then, when the co-occurrence of a cue and an outcome is not reliable, responses should approximate the distribution of co-occurrences. However, if unlearning is an inherent part of learning speech cues, then cues will be downweighted when they do not predict outcomes. Importantly, the downweighting of unreliable (*non-discriminative*) cues, along with strengthening of reliable (*discriminative*) cues, can lead to discriminative cues having greater relative strength, allowing listeners to ignore the unhelpful non-discriminative cues and base their responses on the discriminative ones.

Unlearning has been shown to affect acquisition of semantic categories, specifically, visually presented 'species of alien' objects (Ramscar, Yarlett, et al., 2010). A salient visual cue corresponded to particular labels (outcomes) on most trials, but corresponded to *different* labels on one quarter of trials. This cue was therefore frequent, but not fully discriminative. In order to correctly identify the labels, participants had to learn to ignore the non-discriminative salient cue and use a set of more subtle cues for selecting labels. The critical manipulation was the order of cues and outcomes: either labels preceded cues or cues preceded labels. Results showed that participants did equally well in either presentation order for high-frequency items, but for low-frequency items, accuracy was higher in the discriminative (cue–outcome) order than the non-discriminative (outcome–cue) order. In the non-discriminative order, participants based their responses on the salient but non-discriminative cue. But in the discriminative order participants learned to use the discriminative cues to instead select the correct label. This seems to be because they were able to make use of feedback from prediction error to *downweight the non-discriminative cue*.

As a consequence of this same principle, suffixes have a learning advantage over prefixes due to their position in the sentence (Clair, Monaghan, & Ramscar, 2009; Hoppe, 2016) and Chinese character pronunciation and meanings are learned better with a delay between

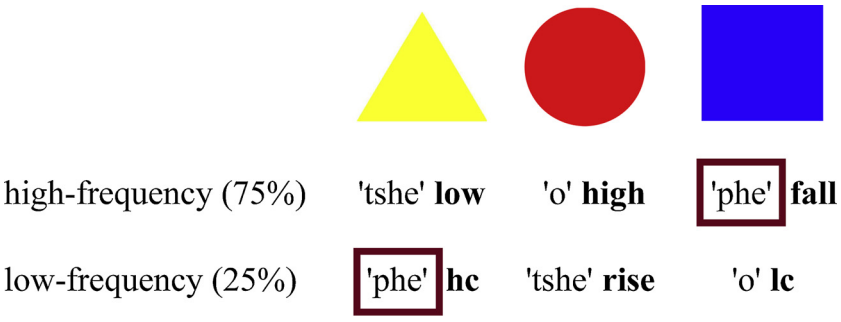


**Fig. 7.** Stimuli in Experiment 2. Outcomes were three coloured shapes (top). Cues occurred either with high frequency (second row) or low frequency (bottom). Tones (in bold) were reliable, *discriminative* cues to the image outcomes (hc = high-checked; lc = low-checked). The same three base syllables occurred in both the high- and low-frequency sets of words; however, critically, each base syllable corresponded to a different image between the high-frequency set and the low-frequency set. For example, the syllable 'phe' occurs with both yellow triangle and blue square. This made the base syllables a *non-discriminative* cue. In order to correctly learn the correspondences, participants had to ignore the syllable and base responses on the tones. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

character presentation and Pinyin or translation compared to simultaneous presentation (Chung, 2003). Similarly, Colunga et al. (2009) showed that rather than a bidirectional mapping between word and referent, word learning is better characterised as a predictive relationship, in which the linguistic cues are used to predict the intended meaning of an utterance.

While Ramscar, Yarlett, et al. (2010) investigated learning of semantic categories, in the present study, the order effect was tested for learning of speech sounds. Experiment 2 tests the prediction that speech cues are learned better when they precede the semantic outcomes. As in Ramscar, Yarlett, et al. (2010), one set of stimuli is simple, without internal structure and the other set is complex and variable with both helpful, discriminative cues and unhelpful non-discriminative cues. Non-discriminative cues are those that have a high probability of occurring with the outcome, but which are unreliable because they also occur with other outcomes. The temporal structure of learning events is manipulated such that for each participant either the simple stimuli or the complex stimuli occur first on each trial. When the complex stimuli occur first, competition can occur between the various cues, leading to an increase in weighting for reliable cues and a decrease for unreliable cues. When the simple stimuli occur first, there is no variation in the cues and thus no cue competition. In this case, learning is expected to reflect the probabilistic relationship with the complex stimuli.

If participants learn speech sounds as a result of cue competition and negative evidence, as predicted by discriminative learning, we expect higher accuracy in the discriminative, compared to the non-discriminative order. If participants learn speech sounds by positive evidence alone, we expect to see no difference between conditions, because the stimulus co-occurrences on each trial are the same in the two conditions.

### 3.1. Method

The online game set up was the same as Experiment 1.

#### 3.1.1. Participants

Participants were 93 native speakers of US English who did not participate in Experiment 1.

#### 3.1.2. Stimuli

Because the present study investigates speech sound acquisition, in what follows, the acoustic stimuli will be referred to as *cues* and the visually presented semantic stimuli will be referred to as *outcomes*. This terminology will apply regardless of the presentation order. Auditory stimuli (*cues*) were three different base syllables ('tshe', 'o', 'phe') with six different lexical tones. Taiwan Southern Min has retained seven of the eight late Middle Chinese tones. The present study includes six of these: high ('yin level', *Yin Ping*), falling ('rising', *Shang Sheng*), low ('yin departing', *Yin Ru*), low-checked ('yin entering', *Yin Ru*), rising ('yang level', *Yang Ping*) and high-checked ('yang entering', *Yang Ru*).

The high and low tones have a flat contour; the rising and falling tones have, as their names suggest, rising and falling contours, respectively; the checked tones begin with an initial flat contour and end with a sharp drop. Two different tones occurred with each base syllable (e.g. 'tshe_low', 'tshe_rising'), resulting in six different tonal syllables. Because more variable stimuli has been shown to facilitate learning through downweighting of non-discriminative acoustic dimensions (Nixon et al., 2016; Rost & McMurray, 2010), two tokens of each tonal syllable were produced for acoustic variability. Visual stimuli (*outcomes*) were three coloured shapes (red circle, yellow triangle, blue square). Each image was randomly assigned to an auditory pair and this mapping was the same for all participants.

#### 3.1.3. Experiment design

During training, three tonal syllables occurred with high frequency (75% of training trials) and the other three with low frequency (25%). Importantly, although the same three base syllables occurred in both high- and low-frequency sets, the syllables corresponded to different image stimuli in the two sets (see Fig. 7). That is, a given base syllable corresponded to one image on 75% of trials and another image on the other 25% of trials. For example, the base syllable (e.g. 'phe') corresponded to a given image in the high frequency stimuli (e.g. 'phe_falling'; blue square), but corresponded to a different image in the low frequency stimuli (e.g. 'phe_high checked'; yellow triangle). Therefore, the base syllable did not reliably predict the target images. To correctly identify images corresponding to low-frequency stimuli, participants needed to ignore the more salient cue – the base syllable – and instead use tone to select the images.

#### 3.1.4. Trial procedure

Experiment 2 consisted of two phases: training and test. During training, there was only one image on the screen and one auditory word was presented. Participants simply clicked the image to continue to the next trial. Stimuli were presented in one of two orders. Either spoken words (cues) were presented, followed by images (discriminative order) or images (outcomes) were presented, followed by spoken words (non-discriminative order). There was an inter-stimulus interval of 1200 ms and an inter-trial interval of 1000 ms (Fig. 9).

During test, one image appeared on each trial with three unlabelled 'key' icons (as on a keyboard) underneath indicating the spoken words. Three auditory stimuli (either the high- or low-frequency set) were presented in random order. Unlike the training phase, there were equal numbers of high- and low-frequency items in the test. The corresponding key icon was highlighted as each auditory stimulus was presented. Participants responded by clicking one of the keys to select the corresponding spoken word.

### 3.2. Analysis and results

A glmer model tested effects of presentation order on likelihood of participants selecting target versus competitor words. The model included a two-level factor of training frequency (high versus low), a two-

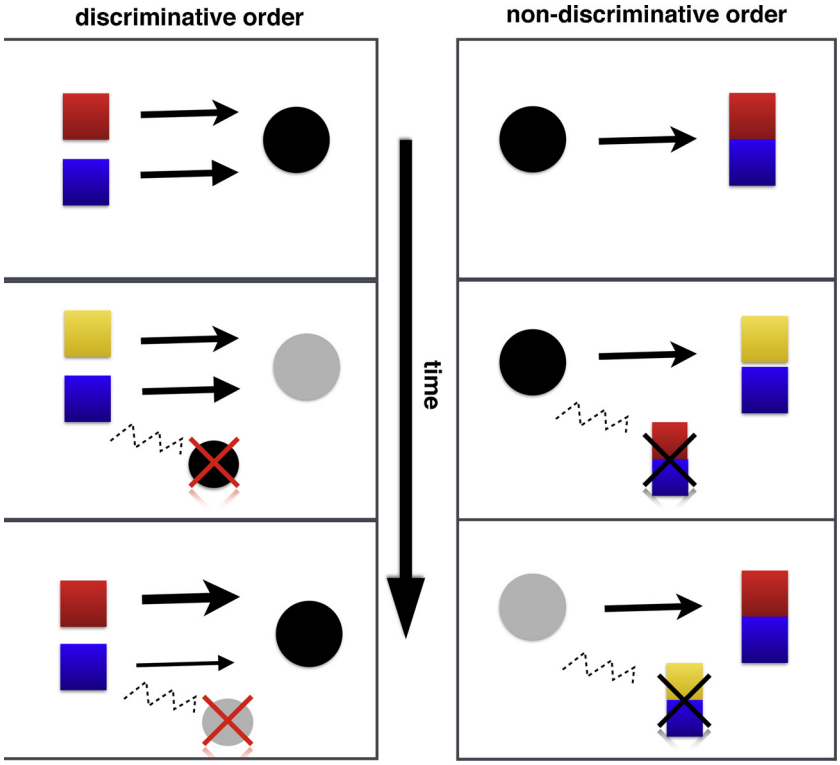


**Fig. 8.** Illustration of unlearning and the effect of cue–outcome order in Experiment 2. In this illustrative example, cues are the colours of the squares: red, blue or yellow; outcomes are the circle colours: black or grey. Arrow thickness indicates (very approximately) connection strength. **Discriminative order (left).** Cues precede outcomes, so they can compete for connection strength. Top panel: red and blue predict the black outcome and gain connection strength. Middle panel: blue and yellow gain connection strength to the grey outcome – simultaneously, they lose connection strength to the black outcome. Bottom panel: in a later trial, blue and red again occur with the black outcome. But the connection weight from blue has weakened (thinner arrow), because it occurred with a different outcome in a previous trial (i.e. in the middle panel). Blue is downweighted so that its connection strength becomes weak. Over time, *blue is unlearned*. In addition, the connection strength of red and blue to the grey outcome weakens, due to their occurrence with the black outcome on the current trial. **Non-discriminative order (right)**. Outcomes precede cues, so there is no competition between cues for connection strength. Top panel: connection strength increases between co-occurring items. Middle and bottom panels: connection strength increases between co-occurring items and decreases between items that do not co-occur. Connection strength simply fluctuates up and down. No difference in the relative connection strength between cues and outcomes occurs. Only the conditional probabilities are learned. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

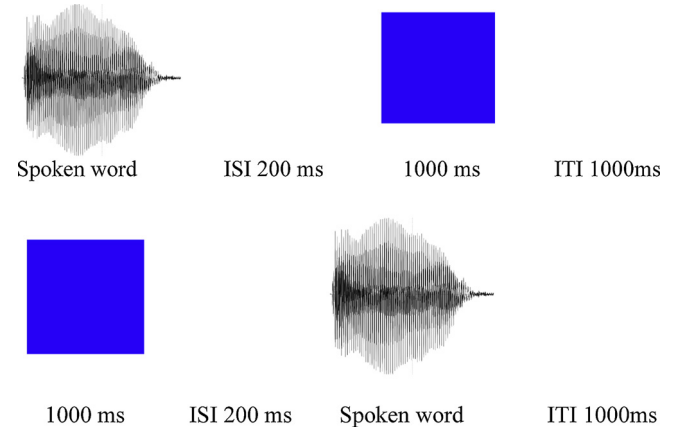


**Fig. 9.** Training trial procedure Experiment 2. In the discriminative order (top), the spoken cues precede the image outcome; in the non-discriminative order (bottom), the image outcome precedes the cues.

level factor of order (discriminative versus non-discriminative) and the two-way interaction, each of which significantly improved model fit. The distractor words (i.e. words in which neither the syllable nor the tone occurred with the image outcomes during training) are not included in the analysis as there were very few distractor responses and they are not relevant to the hypothesis. Participant gender was also tested, but did not improve model fit, so was removed. To account for differences between items, participants and the effect of frequency on participants, random intercepts were included for item and the interaction between participant and frequency. Models with random slopes did not converge and participant random slopes led to high correlations. Based on results of Ramscar, Yarlett, et al. (2010), a significant effect was expected only for low-frequency items.

The model summary is shown in Table 4. The response variable was competitor (0) versus target (1). On the intercept is the discriminative condition for low-frequency items. The interaction between condition and frequency was significant. Most importantly, there were

**Table 4**
Model summary for the effects of cue–outcome order on selection of the target versus competitor word. Model is dummy coded.

| Fixed effects | Estimate | Std. error | $z$-value | $Pr(<\|z\|)$ |
|---|---|---|---|---|
| (Intercept) | 1.0430 | 0.4051 | 2.575 | 0.0100 |
| Condition = non-discriminatve | −1.1984 | 0.5071 | −2.363 | 0.018 |
| Frequency = high | 2.8999 | 0.6139 | 4.723 | < 0.001 |
| Non-discriminative:high | 1.7634 | 0.7753 | 2.275 | 0.023 |

significantly fewer correct target responses in the non-discriminative condition, compared to the discriminative condition for low-frequency items. Accuracy was significantly higher for high-frequency than low-frequency items, but for high-frequency items, there was no difference between conditions ($p = 0.272$).

A visualisation of the model estimates for Experiment 2 is shown in Fig. 10. The estimated proportion of clicks on the correct target word in the cue–outcome order (left) and the non-discriminative, outcome–cue order (right) is shown for high-frequency (blue, dashed line) and low-frequency stimuli (red, solid line). For high-frequency stimuli, the target word was correctly selected in both conditions with very high accuracy and with no difference between conditions. For low-frequency stimuli, participants selected the target significantly more often in the cue–outcome order than the outcome–cue order.

### 3.3. Discussion

Experiment 2 investigated the effects of the temporal order of cues and outcomes. Learning of low-frequency speech cues was significantly better when cues preceded outcomes, compared to when outcomes preceded cues. This demonstrates that there is an asymmetry in the structure of learning events. Rather than a simple 'association', a bi-directional relation, the relationship between acoustic cues and their outcomes appears to be predictive. Learning speech sounds seems to be a process of acquiring expectations from acoustic cues about following semantic outcomes (see also Colunga et al., 2009). When multiple cues

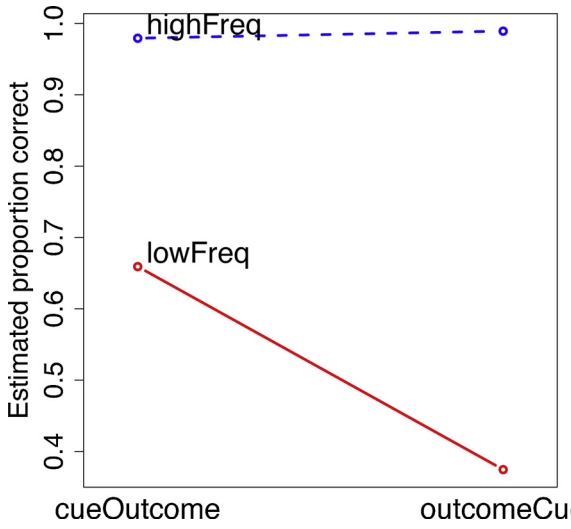


**Fig. 10.** Plot of the model estimates for the results of Experiment 2. Estimated probability of correct responses is on the *y*-axis for high-frequency items (blue, dashed line) and low-frequency items (red, solid line) for discriminative, cue–outcome order (left) and the non-discriminative, outcome–cue order (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

are present and they differ in their informativity about the outcome, cue competition will lead to greater connection strength for informative cues compared to uninformative cues (Ramscar, Yarlett, et al., 2010).

Notably, this is due not only to the increases in connection strength for co-occurring cues and outcomes, but also due to weakening of connections from present cues to absent outcomes. When all the various cues occur first, predictions about the outcomes can be made based on each of the cues. Cues that predict the outcomes are learned (connection weight increases), while cues that do *not* predict the outcome, i.e. are present when an outcome is not present, are unlearned (connection weight decreases). This leads to a difference in cue weighting such that cues that predict outcomes have strong connections and those that do not reliably predict the outcome have weaker connections. This learning of the value of cues may have an effect beyond the specific cue–outcome pair learnt in a specific event. It is this difference in cue–outcome weight that is proposed to develop with experience and lead to speakers' knowledge of – or expectations about – which cues are important and unimportant in their language, as we saw in the introduction. Perception of cues that are valuable for predicting outcomes is honed and perception of those that are not valuable for predicting outcomes becomes poor, both as a consequence of the error-driven adjustments to cue–outcome weight.

A different learning trajectory emerged in the present results when the simple stimuli occurred first. As with the complex stimuli, on some trials one outcome would occur, in which case weight would increase, and on other trials another outcome would occur, in which case the cue weight to the first outcome would decrease. But because the cues in this stimulus were always the same, there was no opportunity for competition between cues, to enhance cue strength to discriminative cues and weaken strength to poor cues. Therefore, the association weight was based on the statistical probability of the stimulus being followed by the outcome. This is what we would expect with, for example, Hebbian models (Hebb, 1949) or models based on co-occurrence statistics. In summary, although the stimuli were identical between conditions, what was learned differed, due to the difference in the predictive structure of the training trials.

## 4. General discussion

Speech acquisition requires honing of discrimination of acoustic cues that are relevant for the particular language being acquired. It also involves reduction of discrimination of cues that are irrelevant. This leads to a system well suited to the native language, but with disadvantages for learning non-native languages. Various accounts of this process have been proposed in the literature. The present study suggests that this process may be at least partially attributable to error-driven, discriminative learning.

Two experiments tested two predictions of error-driven learning derived from the Rescorla–Wagner model, namely *cue competition* and *unlearning*, in the acquisition of non-native speech sounds. Experiment 1 showed that the *blocking* effect, well-known from learning theory (Kamin, 1968, 1969; Rescorla, 1988; Rescorla & Wagner, 1972), also affects acquisition of non-native acoustic cue dimensions. This shows that speech sound acquisition involves cue competition, as predicted by the Rescorla–Wagner equations. Since the statistical distribution of the tested cue was identical in both conditions, the listeners did not learn merely by picking up on distributional statistics of acoustic cues. Instead, learning appears to be driven by uncertainty – or *error* – and the informativity of a given cue for reducing that error (Arnon & Ramscar, 2012; Ramscar, Dye, & McCauley, 2013). If the first cue is already informative for predicting the outcome, the listener lacks the uncertainty necessary to learn additional cues. In short, knowledge of an already-learned acoustic cue can block later learning of new cues. As discussed below, this bears striking resemblance to the pattern of results seen in first and second language acquisition and may help explain effects of so-called 'transfer' from the first language to the second.

In the real world, the speech signal is complex and always contains multiple acoustic cues. This means that cue competition comes into play in multiple ways. In classic blocking experiments, as well as in Experiment 1, a single cue was learned as a strong predictor. Given that the real-world speech signal is highly variable, cues may sometimes spuriously co-occur with a given outcome, but are not consistent or occur with other outcomes, so are not reliable cues. In these cases, as shown in Experiment 2, learning is affected by the predictive structure of learning events, in particular the temporal order of cues versus outcomes. If unreliable speech cues occur before an associated outcome – as in the discriminative, cue–outcome order in Experiment 2 – predictions about the following outcome can be generated, and adjustments made when the outcome differs from expectations. This leads to the unreliable cue being downweighted or *unlearnt*. However, if the semantic outcome occurs before the cue – as in the non-discriminative, outcome–cue order – only the statistical probability of co-occurrence will be learnt. In the experiment, this led to the competitor item being learnt in the non-discriminative order, due to the high correlation between the segmental syllable and the competitor image.

Experiment 2 showed that unlearning plays an important role in the learning of non-native cue dimensions. Learning not only occurs from co-occurrence of cues and outcomes; a vital part of the learning process is learning to ignore unreliable cues. Experiment 2 showed that this unlearning is more likely to occur when the structure of learning events is discriminative – that is, when there is the possibility for cue competition to weaken the connection strength from unreliable cues.

Cue competition was involved in both experiments. However, the role of cue competition differed between experiments. In Experiment 1, there was no cue competition in the pre-training phase. Only one auditory cue was presented, which always occurred with one correct outcome. Only in the following training phase did cue competition come into play. In this case, the first and second cues were competing to predict the semantic outcome. All learning throughout the experiment was 'positive' learning, i.e. strengthening of cue–outcome weights. In the blocking condition, the first cue 'won' the competition, because it had already gained associative strength during the pre-training phase. This meant that for the remaining trials, the increase in strength was small, because the error remaining in the model was small, and hence the second cue could not 'catch up'.

In Experiment 2, in contrast, cue competition was available from the

beginning of the experiment, but only in the discriminative (cue–outcome) order condition. This was because when the complex, variable cues in the acoustic stimuli occurred first in the trial, the unreliable, non-discriminative cues could be downweighted due to cue competition. However, when the outcomes occurred first, there were no alternative cues that could compete. In this case, because no cue competition could occur, learning was instead based on the statistical structure or co-occurrence probabilities. Responses were based on the segmental syllable that occurred most often with a particular image. In the high-frequency stimuli, this was the correct response. In the low-frequency stimuli, this was the competitor image.

### 4.1. Implications for non-native speech acquisition

The present results may provide a way to account for the broad pattern of changes in speech sound acquisition from the first months of life into adulthood discussed in the introduction. When adults learn a second language, they often have influences from their native language, such as an 'accent' in production or difficulties discriminating some sounds that are not relevant for discriminating messages in their native language. From an error-driven learning perspective, these can be explained as cases of both blocking and unlearning. After years, even decades, of experience with native speech cues, listeners have strong expectations about which cues are informative and which are not. Cues that are not discriminative have been downweighted, such that differences are difficult to perceive. This unlearning is useful in the native language where these cues are irrelevant, but if these cues are used in the target language, *reversal* of unlearning is required in order to learn to discriminate them again.

While uninformative cues are weakened, discriminative cues are strengthened. These already-learned discriminative cues may block learning of new cues in the non-native language. This idea is supported by the finding that the degree to which the native language interferes with learning the non-native language, depends on the *relationship* between the specific native and non-native cues (e.g. Best et al., 1988; Flege, 1987, 1995). When the target language uses a sound that is similar (but not identical) to the native language, such as the sounds /u/ and /t/ in English and French, production is less accurate: both English-L1 French-L2 and French-L1 English-L2 speakers produced second formant (in /u/) and voice onset time values (in /t/) that were significantly different to their native speaking counterparts (Flege, 1987). However, when the sound in the target language is not similar to any sound in the native language, such as French /y/ for English speakers, production is more accurate (Flege, 1987). The blocking effects found in Experiment 1 provide a straightforward explanation for this pattern. There is a greater degree of transfer from the native-language when learning similar sounds, compared to dissimilar sounds, because the native sounds are more likely to block learning. The blocking effect may also help explain the recent finding that distributional learning experiments are less effective for adults than infants (Wanrooij, Boersma, & van Zuijen, 2014), as adults' learning is more likely to be blocked by previous learning.

Japanese native listeners often have trouble discriminating English /l/ and /r/ (Goto, 1971; Miyawaki et al., 1975), often relying on the second formant (F2), while English listeners rely on the third formant (F3) (Iverson et al., 2003; Lotto, Sato, & Diehl, 2004; Yamada & Tohkura, 1990). However, when Lim and Holt (2011) trained native Japanese speakers with stimuli with high variability in the F2 dimension, weighting of F2 was weakened and accuracy improved. It seems as if unlearning the non-discriminative cue allowed the listeners to reassign connection strength to the discriminative cue, F3.

In addition, listeners' ability to discriminate non-native speech sounds remains good for sounds that are not used discriminatively in the native language, but which are also *not used non-discriminatively* in the native language. This is the case with Zulu clicks, which are discriminated well by native English infants and adults (Best et al., 1988). It seems that, because English speakers tend to have little to no exposure to unreliable variation of clicks in speech, these sounds are not *unlearned*.

### 4.2. Implications for native-language speech acquisition

While the present study focused on learning of second-language speech sounds, similar principles may also apply in first language acquisition. Learning the meanings of words for colours and numbers poses a challenge for young children, and these words are acquired late, despite their high frequency of occurrence in child directed speech. However, when presented in discriminative order (objects before words, in this case) learning improves (Ramscar, Dye, et al., 2011; Ramscar, Yarlett, et al., 2010). Words are also more learnable when referents are attended before being spoken, compared to when they are not visually attended (Cartmill et al., 2013).

Like these examples of children's word learning, it is possible that early speech sound acquisition is also learned discriminatively. Werker and Tees (1984) noted that honing of infant speech sound perception for the native language occurs at about the time that the first words are being learnt and proposed that these events might be related. This view has been challenged, the argument being that in the first few months of life, infants do not yet have a sufficient lexical knowledge to support discrimination based on lexical contrasts (e.g. Maye & Gerken, 2000). This claim is based on the assumption that minimal pairs would be required. There are at least two possible explanations for how infants could start to build up knowledge of native speech in an error-driven model. Firstly, as shown in Experiment 2, error-driven learning requires learning events to have a discriminative, cue-outcome structure. When learning events do not have predictive structure, this results in patterns of learning that resemble probability tracking. It is possible that infants at this early age, either due to the stage of cognitive development or due to the structure of their environment, are not yet able to learn from prediction error, because they have not yet developed the suitable set of outcomes to discriminate between. However, it is an assumption of discriminative learning models that any discriminable perceptual information can participate in learning. Therefore, a second possibility is that error-driven learning occurs from the beginning, based on all and any perceptual events in the environment. In this case, the temporal nature of speech leads to a situation in which cues that are temporally early in the speech stream may predict temporally later cues. If young infants learn the sounds of their language in this way, they would acquire phonotactic knowledge at the same time as gradually losing discrimination ability for non-discriminative sounds as they get older, as has been documented for infants as they reach the second half year of life (Werker & Tees, 1984). Further research is needed to investigate this possibility.

### 4.3. Relation to learning in other domains

The blocking effects discussed above are also observed for many other aspects of human learning. In language, blocking has been shown to occur during second language morphology and vocabulary acquisition (Arnon & Ramscar, 2012; Chung, 2003; Ellis & Sagarra, 2010). In word learning, if multiple cues occur with an outcome with complete reliability during training, then learning is poorer than when cues are variable (i.e. only occur on 75% of trials). If all cues are completely reliable, learning is blocked. The variability means that certain cues are absent on some trials, allowing learning to increase for the cues that are present (Monaghan, Brand, Frost, & Taylor, 2017).

Furthermore, Apfelbaum and McMurray (2017) show that when multiple word candidates are activated, learning does not wait until word processing is finished and a single candidate is selected. Instead, connection weights are formed between – in their case, visual target – outcomes and these partially activated word candidates in real time. If presentation of the visual targets is delayed so that word activation is

complete before presentation, then these spurious associations are reduced.

In both Ramscar, Yarlett, et al. (2010) and the present study, one set of stimuli was selected to be complex and the other simple. In Ramscar, Yarlett, et al. (2010), the visual semantic categories contained complex cues and labels were considered nondivisible chunks. In the present study, the modalities of the stimuli were reversed. Words contained complex speech cues; images were featureless without internal structure. However, in the real world, both objects and speech are highly complex. During first language acquisition, infants will be exposed to learning events in which either speech or objects occur in discriminative order at different times, allowing them to learn discriminatively about both speech sounds and objects in the world over time.

Finally, it should be noted that the Rescorla–Wagner model has acknowledged shortcomings (reviewed in Miller et al., 1995) and there are aspects of the model still under debate (see also Kapatsinski, 2018, for a discussion and some solutions). For example, in the Rescorla–Wagner model, connection weights are adjusted to outcomes whether the outcomes are present or absent, but weights are only adjusted to cues that are present. There is discussion about whether connection weight adjustments are also necessary for absent cues. Van Hamme and Wasserman (1994) tested whether weights decrease to absent cues in a causal judgment task when a set of three food types (cues) were used to assess the cause of an allergic reaction. In this set up, the food was judged by participants as a less likely cause of the allergy on trials when the food was not present. On the basis of their causal judgment results, Van Hamme and Wasserman (1994) argued for an updated version of the Rescorla–Wagner model in which weights to absent cues are adjusted directly.

Just how to deal with this issue is still under debate. In the Rescorla–Wagner model, absent cues are dealt with indirectly, since cues are competing for limited association strength of the outcome. When present cues increase in strength, they gain relative strength over the cues that are not present, i.e. the cues that are not present do not decrease in strength absolutely, but they lose strength relative to the other cues. In the van Hamme and Wasserman study, there was a limited set of only three cues and one outcome that occurred or did not occur, so it has been argued that the absence of cues may be more salient and therefore may be more likely to be reflected in associative strength under these circumstances. Also, implicit learning seems to work best in automatic, unconscious process (Reber, 1989), but in Van Hamme and Wasserman (1994) participants were making conscious, evaluative judgments about the causal relation. This may increase the likelihood of participants making logical inferences, or perhaps using the allergy response as a cue to predict the cause, essentially reversing the cue–outcome structure. But in any case, despite debate over details of the best implementation of the model, the Rescorla–Wagner model and error-driven, discriminative learning more generally, effectively capture many learning phenomena that are missed by purely statistical learning models.

## 5. Conclusion

In summary, the present study shows that, while experience of linguistic input drives learning of speech sounds, learning does not result directly from perceiving statistical distributions of cues in the environment. Neither does it result only from strengthening connections between co-occurring stimuli. Instead, the combined results of the two experiments suggest that learning is discriminative, driven by uncertainty and prediction error and that weakening of uninformative speech cues plays a crucial role in learning.

A practical conclusion that can be drawn from the present results is that for non-native speech sound acquisition, learning of acoustic cues may be more effective when the cues occur before, rather than after, the semantic outcomes, so that learning can occur from prediction and prediction error. Thus, non-discriminative cues can be downweighted, allowing for better learning of the discriminative cues.

## Acknowledgements

The present experiments were conceived and largely designed while the author was based at the New Zealand Institute of Language, Brain and Behaviour (NZILBB), University of Canterbury. Many thanks to Clay Beckner and Robert Fromont for invaluable technical support and advice, Jen Hay and other members of NZILBB and members of the Quantitative Linguistics Group, University of Tübingen for much valuable discussion during experiment set up, to our speaker for producing the stimuli, to Michael Ramscar, Yu-Ying Chuang and Fabian Tomaschek and to Padraic Monaghan and two additional anonymous reviewers for very helpful suggestions on earlier versions of this manuscript. Also to Harald Baayen for the 'Of mice and men' title suggestion. The roofrunner experimental platform was developed by Chun-Liang Chan and the visuals created by Kayo Takasugi. This project was made possible through the support of a subaward under a grant to Northwestern University from the John Templeton Foundation (Award ID 36617) and an ERC Advanced Grant (Grant number 742545). The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation or the ERC.

## Appendix A

$$\Delta w_{ij}^{(t)} = \begin{cases} \text{(a) } 0 & \text{if } c_i \notin C_t, \\ \text{(b) } \alpha_i \beta_j \left( \lambda - \sum_m \mathrm{I}_{[c_m \in C_t]} w_{mj}^{(t-1)} \right) & \text{if } c_i \in C_t \wedge \ o_j \in O_j, \\ \text{(c) } \alpha_i \beta_j \left( 0 - \sum_m \mathrm{I}_{[c_m \in C_t]} w_{mj}^{(t-1)} \right) & \text{if } c_i \in C_t \wedge \ o_j \notin O_j \wedge o_j \in O_{1,\ldots,t-1}, \\ \text{(d) } 0 & \text{otherwise.} \end{cases} \quad (2)$$

## References

Arppe, A., Hendrix, P., Milin, P., Baayen, R. H., Sering, T., & Shaoul, C. (2015). *ndl: Naive discriminative learning. R package version 0.2.17*. https://CRAN.R-project.org/package=ndl.

Apfelbaum, K. S., & McMurray, B. (2017). Learning during processing: Word learning doesn't wait for word recognition to finish. *Cognitive Science, 41*, 706–747.

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE, 12*, e0174623.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*, 292–305.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*, 438.

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience, 31*, 106–128.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution, 2*, 160–176. https://doi.org/10.1093/jole/lzx001.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.). *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Timonium, MD: York Press.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America, 109*, 775–794. https://doi.org/10.1121/1.1332378.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 345.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 11278–11283.

Cheng, S., & Sabes, P. N. (2007). Calibration of visually guided reaching is driven by error-corrective learning and internal dynamics. *Journal of Neurophysiology, 97*, 3057–3069.

Chung, K. K. (2003). Effects of pinyin and first language words in learning of Chinese characters as a second language. *Journal of Behavioral Education, 12*, 207–223.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*, 804–809.

Colunga, E., Smith, L. B., & Gasser, M. (2009). Correlation versus prediction in children's word learning: Cross-linguistic evidence and simulations. *Language and Cognition, 1*, 197–217.

Cristià, A., McGuire, G. L., Seidl, A., & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics, 39*, 388–402.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. *The Quarterly Journal of Experimental Psychology, 36*, 29–50.

Eimas, P. D. (1985). The perception of speech in early infancy. *Scientific American, 252*, 46–53.

Ellis, N. C., & Sagarra, N. (2010). The bounds of adult language acquisition: Blocking and learned attention. *Studies in Second Language Acquisition, 32*, 553–580.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review, 120*, 751.

Flege, J. E. (1987). The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics, 15*, 47–65.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research,* 233–277.

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /i/ and /l/ accurately. *Language and Speech, 38*, 25–55.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "l" and "r". *Neuropsychologia*.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America, 100*, 1111–1121.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological approach.* Hoboken, NJ: John Wiley & Sons.

Hoppe, D. (2016). *How suffixes enhance the discrimination of linguistic contrasts through their position in sequence (Unpublished masters' thesis).* University of Tübingen.

Houwer, J. D., & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *The Quarterly Journal of Experimental Psychology: Section B, 55*, 289–310.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87*, B47–B57.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. *Miami symposium on the prediction of behavior: Aversive stimulation,* 9–31.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. *Punishment and Aversive Behavior*.

Kapatsinski, V. (2018). *Changing minds changing tools: From learning theory to language acquisition to language change.* MIT Press.

Kopp, B., & Wolff, M. (2000). Brain mechanisms of selective learning: Event-related potentials provide evidence for error-driven learning in humans. *Biological Psychology, 51*, 223–246.

Lim, S.j., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science, 35*, 1390–1405.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America,* 89.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. *From Sound to Sense, 50*, C381–C386.

Maclagan, M., Watson, C. I., Harlow, R., King, J., & Keegan, P. (2009). /u/ fronting and /t/ aspiration in Māori and New Zealand English. *Language Variation and Change, 21*, 175–192.

Magloire, J., & Green, K. P. (1999). A cross-language comparison of speaking rate effects on the production of voice onset time in English and Spanish. *Phonetica, 56*, 158–185.

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. *Proceedings of the 24th annual Boston university conference on language development*.

Maye, J., Weiss, D., & Aslin, R. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science,* 11.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition,* 82.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science, 12*, 369–378.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PLOS ONE, 12*, e0171935.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla–Wagner model. *Psychological Bulletin, 117*, 363.

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics, 18*, 331–340.

Monaghan, P., Brand, J., Frost, R. L. A., & Taylor, G. (2017). Multiple variable cues in the environment promote accurate and robust word learning. *The 39th annual conference of the cognitive science society (CogSci 2017),* 817–822.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 20130299.

Nixon, J. S., & Best, C. T. (2018). Acoustic cue variability affects eye movement behaviour during non-native speech perception. *Proceedings of the 9th international conference on speech prosody* (pp. 493–497).

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language, 90*, 103–125.

Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica, 40*, 1–18.

Ohala, J. J., Hinton, L., & Nichols, J. (1997). Sound symbolism. In: *Proc. 4th Seoul international conference on linguistics [SICOL],* 98–103.

Olejarczuk, P., Kapatsinski, V., & Baayen, R. H. (2018). Distributional learning is error-driven: The role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard,* 4.

Pegg, J. E., & Werker, J. F. (1997). Adult and infant perception of two English phones. *The Journal of the Acoustical Society of America,* 102.

Pépiot, E. (2014). Male and female speech: A study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers. *Speech Prosody 7,* 305–309.

R Core Team (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology, 8*, 51.

Ramscar, M., Dye, M., Gustafson, J. W., & Klein, J. (2013a]). Dual routes to cognitive flexibility: Learning and response-conflict resolution in the dimensional change card sort task. *Child Development, 84*, 1308–1323.

Ramscar, M., Dye, M., & Klein, J. (2013b]). Children value informativity over logic in word learning. *Psychological Science, 24*, 1017–1023.

Ramscar, M., Dye, M., & McCauley, S. M. (2013c]). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language, 89*, 760–793.

Ramscar, M., Dye, M., Popick, H. M., & O'Donnell-McCarthy, F. (2011a]). The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PLoS ONE, 6*, e22501.

Ramscar, M., Matlock, T., & Dye, M. (2010a]). Running down the clock: The role of expectation in our understanding of time and motion. *Language and Cognitive Processes, 25*, 589–615.

Ramscar, M., Suh, E., & Dye, M. (2011b]). For the price of a song: How pitch category learning comes at a cost to absolute frequency representations. *Proceedings of the 33rd annual meeting of the cognitive science society*.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science, 31*, 927–960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010b]). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science, 34*, 909–957.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118*, 219.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66*, 1–5.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist, 43*, 151.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Vol. Eds.), *Classical conditioning II: Current research and theory: Vol. 2,* (pp. 64–99). New-York: Appleton-Century-Crofts.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy, 15*, 608–635.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80*, 1–27.

Schumacher, R. A., Pierrehumbert, J., & Lashell, P. (2014). Reconciling inconsistency in encoded morphological distinctions in an artificial language. *Proceedings of the annual meeting of the cognitive science society*.

Shadmehr, R., Smith, M. A., & Krakauer, J. W. (2010). Error correction, sensory prediction, and adaptation in motor control. *Annual Review of Neuroscience, 33*, 89–108.

Shafaei-Bajestan, E., & Baayen, R. H. (2018). Wide learning for auditory comprehension. *Proceedings of Interspeech 2018, the 19th Annual Conference of the International Speech Communication Association* (pp. 966–970).
Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla–Wagner model. *Psychonomic Bulletin & Review, 3*, 314–321.
Clair, St., Monaghan, M. C., & Ramscar, P. M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science, 33*, 1317–1329.
Sundberg, U., & Lacerda, F. (1999). Voice onset time in speech to infants and adults. *Phonetica, 56*, 186–199.
Terry, J., Ong, J. H., & Escudero, P. (2015). Passive distributional learning of non-native vowel contrasts does not work for all listeners. *ICPhS*.
Tsur, R. (2006). Size-sound symbolism revisited. *Journal of Pragmatics, 38*, 905–924.
Turnbull, R., Seyfarth, S., Hume, E., & Jaeger, T. F. (2018). Nasal place assimilation trades off inferrability of both target and trigger words. *Laboratory Phonology: Journal of the Association for Laboratory Phonology,* 9.
Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*, 127–151.
Wanrooij, K., Boersma, P., & Benders, T. (2015a]). Observed effects of 'distributional learning' may not relate to the number of peaks. a test of 'dispersion' as a confounding factor. *Frontiers in Psychology, 6*, 1341.
Wanrooij, K., Boersma, P., & van Zuijen, T. L. (2014). Distributional vowel training is less effective for adults than for infants. A study using the mismatch response. *PLOS ONE, 9*, e109806.
Wanrooij, K., de Vos, J., & Boersma, P. (2015). Distributional vowel training may not be effective for Dutch adults. *Proceedings of the 18th international congress of phonetic sciences*.
Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63.
Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How do infants become experts at native-speech perception? *Current Directions in Psychological Science, 21*, 221–226.
Wilbur, J. (2015). *A grammar of Pite Saami.* Language Science Press.
Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*, 245.
Yamada, R. A., & Tohkura, Y. (1990). Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese. *First international conference on spoken language processing*.
Yoshida, K. A., Pons, F., Maye, J., & Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy,* 15.