Implicit and explicit processes in phonological concept learning

Elliott Moreton
Department of Linguistics
Smith Building, CB #3155
University of North Carolina
Chapel Hill, NC 27599–3155 U.S.A
moreton@email.unc.edu

Katya Pertsova
Department of Linguistics
Smith Building, CB #3155
University of North Carolina
Chapel Hill, NC 27599–3155 U.S.A
pertsova@email.unc.edu

Version of January 24, 2024.
Comments are very much welcome, and may be addressed to the authors.

Abstract

Non-linguistic pattern learning uses distinct implicit and explicit processes, which differ in behavioral signatures, inductive biases, and proposed model architectures. This study asked whether both processes are available in phonotactic learning in the lab. Five Internet experiments collected generalisation, learning curves, response times, and detailed debriefings from 671 valid participants. Implicit and explicit learners were found in all conditions and experiments. Objective measures of implicit vs. explicit learning were correlated with introspective self-report. Participants spontaneously discovered and named phonetic features. These findings contradict the common (usually tacit) assumption that "artificial-language" participants learn only implicitly. Learning mode also affected inductive bias: Implicit learning improved performance on family-resemblance patterns relative to biconditionals (if-and-only-if, exclusive-or) in two experiments. The direction of this effect is unexpected under many current theories of how implicit and explicit concept learning differ, and is consistent with models of explicit learning which take pattern-irrelevant features into account.

Keywords: phonotactic learning, phonological learning, concept learning, implicit learning, explicit learning, inductive bias, artificial language

1 Introduction

The experimental study of phonological learning has developed rapidly in recent years, providing a new kind of data about the biases that guide learning. As our knowledge has progressed, it has become clear that many experiments, results, and models in phonological learning have close parallels in work on non-linguistic learning (Finley & Badecker 2010; Lai 2012; Moore-Cantwell, Pater, Staubs, Zobel & Sanders 2017; Moreton 2012; Moreton & Pater 2012a,1; Moreton, Pater & Pertsova 2017; Moreton & Pertsova 2016; Pater & Moreton 2012; Pertsova 2012). This creates the opportunity — and the imperative — for systematic comparative study of human inductive learning across domains.

The present study focuses on one particular comparison. We ask whether phonological learning in the lab is like non-linguistic learning in the lab in that learners may use either or both of two distinct processes, one *implicit*, the other *explicit*, which are engaged by different learning situations, have different inductive biases, and different algorithmic architectures.

This study exploits two under-used sources of information. One is detailed analysis of post-experiment debriefing questionnaires in order to collect participants' reports about their own approach to, and experience of, learning the experimental language, and in order to compare that report with objective measures of performance. The other is evaluation, not just of end-state performance, but of how performance changes over time (learning curves), in order to compare it with the predictions of different learning models. Experiments 1 and 2 focus on identifying correlates of implicit vs. explicit learning modes using single-feature assertions ("Type I" patterns, in the terminology of Shepard, Hovland & Jenkins 1961). Experiments 3, 4, and 5 ask whether the two modes have different inductive biases, by comparing their success in acquiring two-feature if-and-only-if ("Type II") and three-feature family-resemblance ("Type IV") patterns.

This study goes beyond previous work on phonotactic learning by testing the two-systems hypothesis. It goes beyond previous work on the two-systems hypothesis in non-linguistic learning by applying that hypothesis to complex phonological stimuli to facilitate cross-domain comparison. It goes beyond both by uniting so many indices of learning mode in a single study.

This paper is aimed at two audiences simultaneously. One is phonologists who know about phonological learning and are interested in how it relates to learning in other domains, or in how participants approach phonological tasks. The other is cognitive scientists who know about concept learning and are interested in how it relates to learning phonological patterns, which inhabit a much more complex stimulus space than is typically studied.

2 Implicit and explicit concept learning

Studies of inductive learning of featurally-defined non-linguistic patterns (also called "concepts" or "categories"; e.g., "blue and triangular") have led many psychologists to hypothesise two concurrent learning processes, which here we will call the *explicit system* and the *implicit system* (Ashby, Alfonso-Reese, Turken & Waldron, 1998; Kellogg, 1982; Love, 2002; Maddox & Ashby, 2004; Smith, Berg, Cook, Murphy, Crossley, Boomer, Spiering, Beran, Church, Ashby & Grace, 2012; Smith, Zakrzewski, Herberger, Boomer, Roeder, Ashby & Church, 2015). The two systems correspond approximately to the familiar notions of *reasoning* and *intuition*. Each is characterised by a set of putatively co-occurring properties. Several variants of this two-systems hypothesis exist; for critical reviews see Evans (2008); Keren & Schul (2009); Newell, Dunn & Kalish (2011); Osman (2004).[1]

The explicit system is hypothesised to be effortful, conscious, and demanding of attention and working memory. It is proposed to have a "rule-based" architecture, i.e., it can be modelled as serial testing of verbalizable hypotheses; hence, learning is abrupt (as one hypothesis ousts another; Bower & Trabasso 1964), open to introspection, and subject to inductive biases which make it better for patterns which depend on fewer features. Proposals differ as to how the fewer-relevant-features bias arises out of the learning model; e.g., RULEX (Nosofsky, Palmeri & McKinley, 1994b) serially tests candidate rules in order of increasing feature count, whereas the mental-model model (Goodwin & Johnson-Laird, 2013) begins with a set of parochial rules for each instance, which it then progressively amalgamates by detecting and eliminating irrelevant features. Another conjectured source is a preference for rules that are shorter when expressed in natural language (Ciborowski & Cole, 1973; Greer, 1979; Maddox, Filoteo & Lauritzen, 2007; Shepard *et al.*, 1961).[2]. The process generating the bias will become relevant in a post-hoc analysis (Section 9); until then, we will simply hypothesise that fewer relevant features mean faster and more accurate explicit learning.

The implicit system, by contrast, is proposed to be effortless, unconscious, and undemanding of attention or working memory. Architecturally it is proposed to be "cue-based", i.e., the learning model involves incremental weight update on an array of property detectors which are functionally analogous to weighted constraints in linguistic theory (Ashby, Paul & Maddox, 2011; Gluck & Bower, 1988; Nosofsky *et al.*, 1994b; Rescorla & Wagner, 1972).[3] Hence, learning is gradual rather than abrupt, closed to conscious introspection,

[1] Another approach divides knowledge into declarative vs. procedural, but that is not straightforwardly applicable to phonotatic knowledge, which can be implicit without being procedural. We thus focus on the implicit-explicit distinction, as defined by (e.g.) Lee (1995); Mathews, Buss, Stanley, Blanchard-Fields, Cho & Druhan (1989); Reber (1993); Smith *et al.* (2015), rather than procedural vs. declarative.

[2] There are other proposals which posit a bias towards rules that are shorter or otherwise simpler when stated in a model-internal rule syntax (e.g., Feldman 2000,0; Goodman, Tenenbaum, Feldman & Griffiths 2008; Shepard *et al.* 1961; Thaker, Tenenbaum & Gershman 2017). These proposals leave open whether the model is meant to describe explicit processes, implicit processes, or a single-system alternative to the two-systems hypothesis, and so are not further pursued in this paper.

[3] The analogy is looser for exemplar- or cluster-based models, in which stimuli are simultaneously memorised and simplified by adjusting attentional weights to emphasise some features and de-emphasise others, resulting in a population of complex

and faster for patterns which are supported by multiple overlapping cues than for those that are supported by a small number of disjoint cues.

Each system is associated with a distinct syndrome of predicted behavioral effects. Since the explicit system is conscious and effortful, participants are predicted to be aware of whether they are using it or not. Since the end product of explicit learning is an explicit rule that governs the learner's classification responses, explicit learners should show a tight link between classification performance and ability to accurately verbalise the target rule. In an experiment where a partly-correct rule is no help, explicit learners are predicted to fall into two groups at the end of training: those who achieve a high level of classification accuracy and are able to accurately verbalise the target rule, and those who are near chance and state an inaccurate rule or no rule. If trial-by-trial responses are collected during training, an abrupt jump from near-chance to near-perfect performance, and from slow to fast reaction times, might coincide with the discovery of the correct rule and the participant's transition from rule-seeking to rule-using.

In the hypothesised implicit system, on the other hand, the product of learning is a set of continuous-valued weights on an array of property detectors; hence, implicit learning should not facilitate accurate verbalisation of the target rule. Since the weights are updated incrementally and automatically, and since responses are smoothly related to the weights, changes in response probabilities and reaction times should be gradual over time and similar across participants.

The dependence of the explicit system on working memory is hypothesised to bias it in favor of rules that involve simple relations between a small number of features, such as two-feature biconditionals (if-and-only-if and exclusive-or patterns, e.g., "either green or square, but not both"), whereas the parallelism of the implicit system facilitates detection of patterns which are supported by multiple overlapping cues, such as multi-feature family-resemblance patterns (e.g., "differs by at most one feature value from a small green square"). Evidence for the occurrence of these symptoms in non-linguistic learning is summarised in Table 1.

Different experimental conditions facilitate the use of one or the other learning mode. Corrective feedback, instructions to seek a rule, and easily-verbalizable stimulus features elicit more behavioral signatures of explicit learning, while training without feedback, instructions that do not mention rules, and features that are hard to verbalise favor implicit learning (Table 2).

Each system is proposed to be domain-general, i.e., to apply to any concept regardless of the real-world features which define it. The concepts "blue and triangular", "feverish and sniffly", "furry and oviparous", etc. are all grist for the same two mills. Though the verbalizability of the features, or the perceptual

multi-feature property detectors on which responses are based (Anderson, 1991; Kruschke, 1992; Love, Medin & Gureckis, 2004). The gradual updates apply to both the attentional weights and the association strengths between the exemplars and the response category.

Table 1: Behavioral signatures of explicit vs. implicit learning in experiments on non-linguistic learning.

Symptom	Explicit	Implicit	
Report rule seeking/finding/use	yes	no	Bruner, Goodnow & Austin (1956); Ciborowski & Cole (1972)
Can state correct rule	yes	no	Ciborowski & Cole (1973)
Correctness of stated rule predicts performance	yes	no	Lindahl (1964)
Shape of learning curve	abrupt	gradual	Smith, Minda & Washburn (2004)
Progression of RTs	abrupt	gradual	Haider & Rose (2007)
Distribution of test-phase performance	bimodal	unimodal	Kurtz, Levering, Stanton, Romero & Morris (2013)
Structural bias	IFF/XOR easier than family-resemblance	IFF/XOR advantage reduced or reversed	Kurtz *et al.* (2013); Love (2002); Rabi & Minda (2016)

Table 2: Conditions favoring explicit vs. implicit learning in experiments on non-linguistic learning.

	Favors		
Condition	Explicit	Implicit	
Training	with feedback	no feedback	Love (2002)
Instructions	urge rule-seeking	don't mention rules	Kurtz *et al.* (2013); Lewandowsky (2011); Love (2002); Love & Markman (2003)
Intent	intentional	incidental	Love (2002)
Features	verbalizable	not verbalizable	Kurtz *et al.* (2013); Nosofsky & Palmeri (1996)

separability of their physical instantiations, might affect learning (Kurtz *et al.*, 2013; Minda, Desroches & Church, 2008; Nosofsky & Palmeri, 1996; Zettersten & Lupyan, 2020), the processes themselves are proposed to be general-purpose. It follows that both processes ought to be applicable to language, and indeed, both implicit and explicit processes have been found to be involved in language learning (Ellis 1994; for reviews see Lichtman 2013; Rebuschat 2013).[4] A widespread view is that child L1 learning is implicit and domain-specific, while adults learning L2 rely on explicit domain-general problem-solving abilities (Bley-

[4] By "explicit learning", we mean here explicit *inductive* learning, not explicit *instructed* learning where the language learner is told outright what the pattern is.

Vrooman, 1990; DeKeyser, 2003; Paradis, 2004). This is an oversimplification, as there is evidence of implicit morphosyntactic grammar learning in both naturalistic (non-classroom) L2 acquisition (Green & Hecht, 1992; Krashen, 1982) and in artificial-language experiments (Lichtman, 2012; Reber, 1993).

There has been little, if any, study contrasting implicit vs. explicit learning of natural first- or second-language phonotactics.[5] Studies of phonological learning in artificial languages are mainly aimed at explaining natural-language typology, and therefore assume — usually tacitly — that all participants use a single implicit inductive learning process, identical to the one that underpins natural language acquisition and shapes natural-language typology. Criticisms of "artificial-language" methodology as contaminated by explicit learning (e.g., Zhang & Lai 2010) have not presented evidence that it actually is so contaminated. Experimenters may design their experiments to minimise explicit learning (e.g., Do, Zsiga & Havenhill 2016; Glewwe 2019), or exclude data from participants who correctly verbalise the pattern (e.g., Chen 2020; Lin 2023; Moreton 2012; Zellers, Post & Williams 2011), but, with some recent exceptions (Chen 2021; Kimper 2016; Moreton & Pertsova 2016; Moreton, Prickett, Pertsova, Fennell, Pater & Sanders 2021), they rarely analyze implicit and explicit learners separately, nor distinguish wholly implicit learners from failed explicit learners.

Lack of knowledge about the learning-mode variety of phonological learning is an obstacle to progress. Despite their growing importance to phonological theory, we do not know what artificial-language experiments are "about". Are participants really all applying the same processes as each other? Are they applying the same processes as natural L1 or L2 learners? Are there experimental manipulations that encourage the kind of learning the experimenters want to study? Are there ways to distinguish different kinds of learners in the analysis? Do differences in how participants learn lead to differences in what kinds of pattern they learn better?

This study asks whether the inductive learning of phonotactics in the lab is served by implicit and explicit processes like the ones proposed for non-linguistic inductive concept learning. The research strategy is simple: using phonological patterns rather than non-linguistic ones, to vary the conditions in Table 2, observe the effects on the symptoms in Table 1, and compare the results to the predictions of the two-system model.

2.1 Approaches to implicit vs. explicit learning in related areas

This study, motivated by the parallels between the concept-learning literature in psychology and the phonotactic-learning literature in phonology, focuses on the empirical area where those parallels are strongest, namely, experiments in which adult participants classify stimuli on the basis of a featurally-defined pattern. There

[5] There is a sizable literature on instructed vs. naturalistic acquisition of second-language *phonetics*, reviewed in Thomson & Derwing (2015).

are two other neighboring areas in which debate is ongoing as to the relative contributions of implicit and explicit knowledge, and which have been studied in connection with non-linguistic analogues.

One is the learning of phonologically unpatterned wordlike chunks from speech-stream segmentation ("statistical learning", e.g., Saffran, Newport & Aslin 1996). Participants are exposed to an uninterrupted stream of concatenated pseudo-words sampled with repetition from a fixed set, and are then tested on their ability to recognise the pseudo-words in isolation and distinguish them from foils. The statistical dependencies that make it possible to parse out the pseudo-words are phonologically arbitrary dependencies between specific syllables or segments (e.g., Newport & Aslin 2004). Here, both implicit and explicit processes seem to contribute something non-negligible (Batterink, Paller & Reber, 2019; Batterink, Reber, Neville & Paller, 2015). Analogous visual experiments have found predominantly implicit learning (Kim, Seitz, Feenstra & Shams, 2009), but with some role for deliberate attention (Turk-Browne, Isola, Scholl & Treat, 2008).

Another approach involves speech errors occurring during speeded production of sequences of syllables (Dell, Adams & Meyer, 2000). The oft-replicated finding is that a consonant which is restricted by the experimental pattern to a specific syllable position (onset vs. coda) stays in that position when moved to a different syllable by an error more often than a consonant whose position is not so restricted (Anderson & Dell, 2018). Participants usually show signs of implicit learning, such as insensitivity to instructions that reveal the pattern and inability to report it afterwards (reviewed in Dell, Kelley, Hwang & Bian 2021), but that is not always the case (Taylor & Houghton, 2005, Experiment 1). Recently, Smalle, Muylle, Szmalec & Duyck (2017) and Muylle, Smalle & Hartsuiker (2021) have found that children's and older adults' errors on position-restricted consonants are like those of younger adults, but that, unlike younger adults, children and older adults have no tendency to preserve the syllable positions of unrestricted consonants in errors. Muylle *et al.* (2021) speculate that this might be because younger adults have better explicit cognition than the other two groups, whereas implicit learning ability is constant across the lifespan. In button-pressing analogues in which fingers played the role of consonants and thumbs those of vowels, it was found that, although errors respected position in the syllable-analogues as in the language versions, there was no tendency for unrestricted consonant-analogues to preserve position (Anderson & Dell, 2018; Rebei, Anderson & Dell, 2019) — i.e., younger adults in the button-pressing task behaved like children and older adults in the speech task. If Muylle *et al.* (2021)'s speculation is correct, that could mean that the button-pushing task is entirely implicit, whereas the speech task engages some explicit processing in younger adults.

How concept learning, statistical learning, and production learning are related to each other and to natural first- or second language acquisition will be a difficult knot to unpick. One approach would be to compare how the same pattern is learned across all three experimental paradigms, and that can hardly be done without clarifying the role of implicit and explicit processes in each one.

3 Experiment 1

In Experiment 1, the conditions in Table 2 were varied to see if they had the effects in Table 1.[6] The Implicit-Promoting condition was based on a common paradigm in which participants are familiarised using only pattern-conforming instances, then tested on their ability to choose a novel pattern-conforming item when paired with a non-conforming foil (e.g., Carpenter 2006,1,1; Gerken, Quam & Goffman 2019; Greenwood 2016; Kuo 2009; Lai 2015; Moreton 2008,1; Moreton *et al.* 2017; Skoruppa & Peperkamp 2011). The Explicit-Promoting condition differed in that training trials consisted of choosing the conforming member of a conforming-non-conforming pair, a condition which encourages explicit learning in non-linguistic experiments because it asks for explicit judgements and provides explicit corrective feedback (see Section 2).

In the above-cited experiments corresponding to the Implicit-Promoting condition (Carpenter 2006, etc.), the familiarisation task was explained to participants as listening to "words" in a "language", and the test task as distinguishing novel words of the language from nonwords. Using that task here would have meant familiarizing our Explicit-Promoting participants by training them to choose words over nonwords, a task which has no analogue in natural language learning. To improve ecological validity in the Explicit-Promoting condition, participants in both conditions of Experiment 1 were instead told that they would be learning to distinguish words of the target gender from words of another gender. Many natural languages assign gender at least partly on the basis of arbitrary phonological properties (Corbett, 1991, 51–62), and guessing the gender of a new word is something that speakers of such languages must sometimes do in real life, making use of phonological cues among others (Franco, Zenner & Speelman, 2018; Onysko, Callies & Ogiermann, 2013; Zubin & Köpcke, 1984). Each participant's "language" assigned nouns feminine or masculine gender based on a visual or phonological feature chosen randomly from a larger set. Participants were trained, tested, and then given a post-experiment debriefing questionnaire.

3.1 Methods

3.1.1 Stimuli

The audio stimuli (fictitious nouns) were American English nonwords with the prosodic shapes [(əC)VCəC] and [VCəC(əC)]. Main stress fell on the first or second syllable; other syllables' vowels were reduced to [ə]. [7] The stressed vowel was one of [i ɪ e ɛ u ʊ o ɔ]. The consonants were one of [p b t d f v s z]. The schema is

[6] Parts of this section were previously presented as Experiment 1 of Moreton & Pertsova 2016.

[7] Vowel-initial words were used for forward compatibility with other planned experiments. Similar patterns are attested in a number of natural languages; e.g., in Urama (Papua New Guinea; Trans-New Guinea) all verb roots begin with a vowel (Brown, 2009; Brown, Muir, Craig & Anea, 2016). In Ẹdo (also called Bini) and its relative Urhobo (both Nigeria; Niger-Congo, Edoid), all nouns, or nearly all, begin with a vowel (Ọmọruyi, 1986; Kelly, 1969). In Èwùlù (Nigeria; Niger-Congo, Igboid), all noun stems begin either with a vowel or with a syllabic nasal (Utulu, 2020). In Arrernde (Arrernte, Aranda; Australia, Pama-Nyungan), all words are vowel-initial at the surface level (Breen & Pensalfini, 1999).

shown in Table 3. Examples are shown in Figure 1.

Consonants

	Lab		Cor	
voiced	−	+	−	+
−cont	p	b	t	d
+cont	f	v	s	z

Stressed vowels

	−back		+back	
tense	+	−	+	−
+high	i	ɪ	u	ʊ
−high	e	ɛ	o	ɔ

Prosodic shapes

	Disyllabic	Trisyllabic
$\acute{\sigma} = \sigma_1$	*VC*ə*C*	*VC*ə*C*ə*C*
$\acute{\sigma} = \sigma_2$	ə*CVC*	ə*CVC*ə*C*

Table 3: Schema used to construct the auditory nonword stimuli for all experiments.

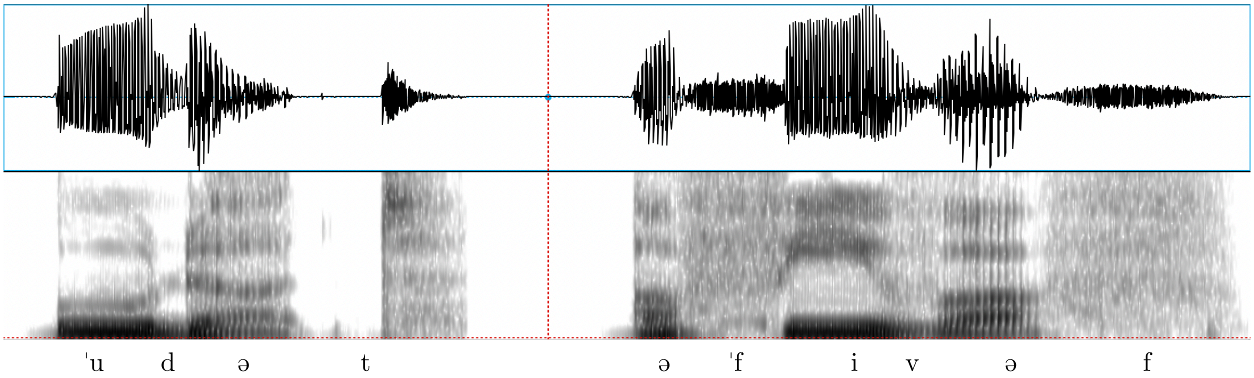


Figure 1: Examples of audio stimuli, shown in Praat (Boersma & Weenink, 2021). Dimensions: 2.21 s × ±0.93 uncalibrated units (waveform) or 5000 Hz (spectrogram).

Six phonological variables were chosen based on the authors' expectations that each would be individually highly salient, i.e., would result in high learning performance in a Type I pattern. Three were chosen with the expectation that they would be easy for linguistically-naïve participants to verbalise: two vs. three syllables, first- vs. second-syllable stress, all consonants different vs. all consonants identical. The other three were chosen with the expectation that they would be hard to verbalise: stressed vowel is front (and unrounded) vs. stressed vowel is back (and rounded), all consonants are fricatives vs. all consonants are stops, and all consonants are labial vs. all consonants are coronal. The reason for making *all* consonants share the property was to make the rule findable regardless of which consonant position or positions the participant happened to focus their attention on. The six variables were crossed to create 64 cells, each of which was filled with 8 randomly-generated nonwords to create a pool of 512 nonwords. We will refer to these variables as "features" henceforth, using the word in its everyday sense rather than in the technical sense of an element in a theory of distinctive features (Jakobson, Fant & Halle, 1952).

Each stimulus was recorded in isolation by a male native speaker of American English from the Upper Midwest at a 44.1 kHz sampling rate. Using Praat (Boersma & Weenink, 2013), they were high-pass filtered with a 10-Hz rolloff at 100 Hz to remove low-frequency noise, and normalised to have the same peak amplitude. The resulting high-resolution WAV-format files were lossily compressed to MP3 and Ogg Vorbis format for use in the actual experiment. The pictures were collected from public-domain sources found on the World Wide Web. Each depicted a familiar object on a white background.

3.1.2 Participants and procedure

Participants were recruited for a study on learning grammatical gender in an artificial language using Amazon Mechanical Turk (Sprouse, 2011). A total of 211 participants completed the experiment. Of these, 20 were excluded from analysis (5 reported a non-English L1, 7 reported taking written notes, 6 reported choosing test-phase responses that were maximally *un*like what they were trained on, 2 fell below the minimum performance criterion of at least 10 correct answers out of 32 in the test phase,[8] leaving 191 valid participants. In addition to the six phonological-feature conditions described above, there were also three visual-feature conditions which will not be discussed here (but see Pertsova & Becker 2020 for some discussion). That left 137 valid participants in the phonological conditions (63 Explicit-Promoting and 74 Implicit-Promoting). No participant, in this or any other experiment, participated in more than one of the experiments reported in this paper.[9]

The experiment was preceded by a sound check, in which potential participants were asked to listen to a single word and type it. Those who were unable to hear the audio were asked not to participate further. Participants were then randomly assigned to one of 24 groups defined by crossing Training Group with Critical Feature and Target Gender.

A unique "language" was randomly generated for each participant, consisting of 128 word-picture pairs, randomly divided into 32 conforming and 32 non-conforming items for the training phase, and another 32 and 32 for the test phase. Grammatical gender was explained as follows:

> This artificial language is like Spanish or French in that it has *grammatical gender*: All nouns are grammatically either feminine or masculine, even if they refer to things like clouds or sidewalks that have no biological sex.

Participants in the Implicit-Promoting group were instructed that all of the words they were to learn would belong to the Target Gender. On each training trial, the participant saw a picture, captioned with its English name, with a button below it (Figure 2, left panel). Mousing over the button played the correct word for that picture in the artificial "language". Clicking the button triggered the next trial after a 250-ms

[8] The criterion was meant to exclude participants who misunderstood the task and systematically attempted to choose the *non*-conforming item. The reason for setting the criterion at 10 incorrect answers out of 32 is that a participant who scored at that level or below was significantly *below* chance performance (two-sided binomial 95% confidence interval).

[9] Mechanical Turk batch sizes for each experiment were chosen with the aim of having at least 12 valid participants for each unique combination of Type (I, II, and/or IV, depending on experiment), Training Condition (Implicit-Promoting or Explicit-Promoting), and assignment of physical dimensions to logical dimensions (6 levels for Experiment 1, 3 for the other experiments). There were 54 such cells of this sort across the 5 experiments. Actual yield per cell varied, with a mean of 11.2 and a standard deviation of 3.8.

delay. All 32 pattern-conforming stimuli were presented in random order, then again in a different random order, and so on until they had been presented four times over. The random order was constrained to consist of four-trial blocks such that each trial within a block came from a different one of the four bins that corresponded to pattern-conforming feature values.

Participants in the Explicit-Promoting group were instructed that they would learn to tell whether a word belonged to the Target Gender by trial and error; and there were systematic differences between the feminine and masculine words which were reliable guides to the right answer. On each training trial, participants saw two pictures, each with a button below it which played the correct word when moused over (Figure 2, right panel). The task was to choose the picture-word pair that had the Target Gender. The response was followed, after 500 ms, by feedback. For a correct response this was the sound of a desk bell. One second after the onset of the bell, the correct response was played again, and two seconds after the onset of that stimulus, the next trial began. Following an incorrect response, the feedback was a sad two-note sequence played on a trumpet, after which the software waited for the participant to click on the correct button before proceeding to the next trial. After all 32 conforming-nonconforming pairs had been presented, they were re-paired, reordered, and re-presented, until they had been presented four times ("timed out"), or until the participant had responded 100% correctly on four consecutive 4-trial blocks ("reached criterion").

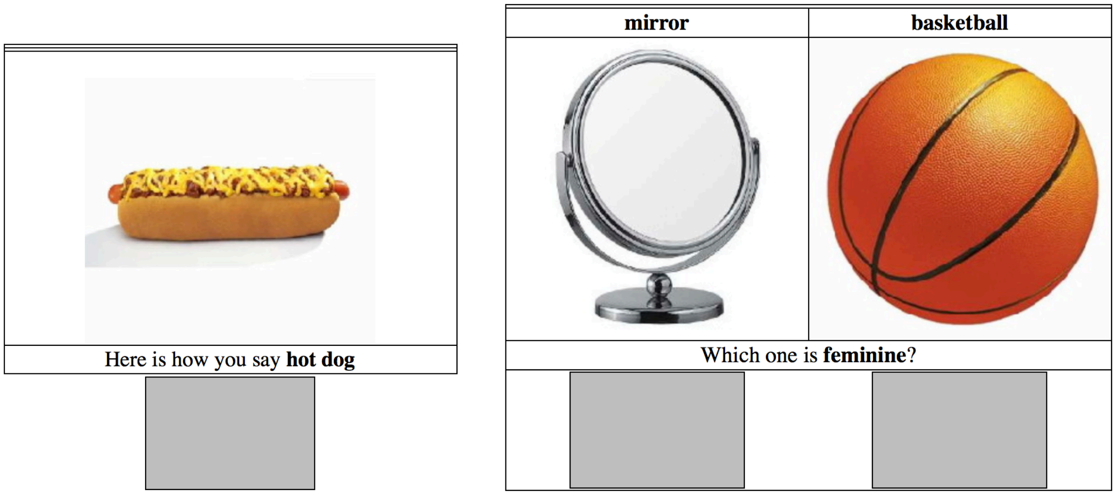


Figure 2: Participant view of a trial in Experiment 1. Left panel: Training phase, Implicit-Promoting condition. Right panel: Training phase, Explicit-Promoting condition, and test phase, both conditions.

Participants in both conditions were instructed to pronounce the audio stimuli aloud before responding. A timestamp was recorded by the server when a trial was transmitted to the participant, and another when the server was notified that the trial had ended, using the `time` function in the `Time::HiRes` module in Perl (Wegscheid, Schertler & Hietaniemi, 2015). Since response times were measured at the server, they

1. How did you approach the learning task (the first part of the experiment? Please choose all that apply: □ Went by intuition or gut feeling. □ Tried to memorize the words. □ Tried to find a rule or pattern. □ Took notes
2. Please describe what you did in as much detail as possible. If you looked for a rule, what rules did you try?
3. How did you approach the test (the second part of the experiment)? Please choose all that apply: □ Chose words that sounded *similar* to the words I'd studied. □ Chose words that sounded *different* from the words I'd studied. □ Chose words that fit a rule or pattern.
4. Again, please describe what you did in as much detail as you can. If you used a rule, what was it?
5. What percent of the test questions do you think you got right?
6. Did you have an ”Aha!” moment, where you suddenly realized what the pattern was? (TRUE/FALSE)
7. If so, please describe the "aha!" moment. When did it happen? What was it you suddenly realized?

Table 4: Post-experiment debriefing questions (1–5: all experiments; 6 and 7: Experiment 2).

include transmission time to and from the participant's computer, as well as the time required to render the page and play the sound files, which add variability to the durations (Høiland-Jørgensen, Ahlgren, Hurtig & Brunstrom, 2016).

The last training trial was followed by the test-phase instructions, identical for both Training Groups. The procedure was identical to the training phase of the Explicit-Promoting group, except that the novel pattern-conforming and non-conforming test items were used, and there was no feedback; either response was followed, after 250 ms, by the next trial. Each of 32 conforming-nonconforming test pairs was presented once.

The experiment was followed by a debriefing questionnaire. In addition to questions about age, gender, and linguistic background, the questionnaire asked the participant to introspect about the learning process and the outcome of learning. The questions asked are shown in Table 4.

3.1.3 Questionnaire coding

Self-report can be used in many different ways to assess implicit vs. explicit learning (Tunney & Shanks, 2003), but there is no way to cleanly divide participants into one group that used exclusively implicit processes, and another that used exclusively explicit ones, because of the possibility of inaccurate self-report and the probability that many participants use some of each. We can only sort participants into more- and less-explicit groups, i.e., groups that are likely to contain a higher or lower proportion of people who relied more on explicit or more on implicit processes. Questionnaire responses were coded according to the following criteria:

Feature stating: Did any of the answers mention any of the *critical* phonological features of the target rule by description (rather than by, e.g., listing letters)?

Rule stating: Did any of the answers state an explicit property of the audio or visual stimulus, and say or imply that the participant's training or test responses were guided by it at any point in the experiment? (Rules that the participant said they tried and abandoned were included when scoring rule-stating.)

Rule correctness: Did the participant report the correct rule? If not, did they report an approximation, a rule that was more than 50% correct? (Rules that the participant said they tried and abandoned were not included in scoring rule correctness.)

Listing: Did any of the answers list sounds, syllables, or letters?

The answers to the free-response questions (Questions 2 and 4) were merged into a single answer for scoring. This was necessary because participants often answered each question, at least partly, in the other question's response box.

Participants' answers to the free-response questions were coded by two of the experimenters using software custom written by Josh Fennell. To minimise criterion drift across experiments, the questionnaires from all of the experiments reported in this paper were coded together, with individual participants' questionnaires occurring in random order so that questionnaires from different experiments were intermixed. Since the only unstressed vowel was schwa, there was no principled distinction between specifying stress location in terms of where schwa was found, and specifying it in by listing the vowel sounds that appeared in a particular position; hence, both response types were arbitrarily scored as feature-stating rather than letter-listing.

Cohen's κ statistic for inter-rater reliability was calculated using the *kappa2* function of the *irr* package in R (Gamer, Lemon, Fellows & Singh, 2019). All of the κs were above 0.8, a level which is typically regarded as indicating high reliability (Cohen, 1960; Landis & Koch, 1977; McHugh, 2012; Munoz & Bangdiwala, 1997).

3.2 Hypotheses and planned analyses

If the explicit system is in fact open to conscious introspection and under voluntary control, then questionnaire responses about the use of that system should reflect performance of its users in the training and testing phases with better-than-chance accuracy. In order to make concrete predictions, participants were classified based on their scored questionnaire responses according to the following schema:

Rule-Seeker: Checked box “Tried to find a rule or pattern” with reference to the training phase.

Rule-Stater: In at least one of their free-response responses, stated a rule. Subdivided into *Correct Rule-Staters*, *Approximately-Correct Rule-Staters*, and *Incorrect Rule-Staters* as scored (see Section 3.1.3).

Memoriser: Checked box “Tried to memorise the words” with reference to the training phase.

Intuiter: Checked box “Went by intuition or gut feeling” with reference to the training phase.

In training conditions where feedback was given, the training phase yields a learning curve, on the basis of which participants were additionally classified according to whether they met the stopping criterion:

Solver: In a condition with feedback, someone who met criterion (16 consecutive correct trials).

A participant who reported using multiple approaches was coded TRUE for each of the relevant categories.

If use of the explicit vs. implicit system is facilitated by the same factors as in visual pattern learning, then more Explicit-Promoting than Implicit-Promoting participants should be Rule-Seekers and Rule-Staters (*Hypothesis 1*).

If a participant states a correct explicit rule, that rule is likely to be the source of their test-phase responses: Correct Rule-Staters should perform near 100%. Participants who did not state a correct rule — the Non-Staters, Incorrect Staters, and Approximate Staters — may be a more heterogeneous group. Their responses could be based on an approximately-correct explicit rule, an outright wrong explicit rule, an implicitly-learned intuition about the pattern, similarity to memorised training stimuli, or even a correct explicit rule that they omitted to state on the questionnaire. Hence, Non-Staters should show a wide distribution of somewhat above-chance performance, and Correct Staters should outperform Approximate Staters (*Hypothesis 2*).

By comparing Solvers with each other, we can compare participants who achieved high performance by different routes to see if differences in the learning curve correspond to differences in self-report. A participant who becomes a Solver by serial hypothesis-testing alone would show near-chance performance until finding the correct rule, whereupon performance would improve to near-perfection and stay there. Once the correct rule is found, the participant can respond to a trial after hearing just one of the two stimuli. Hence, among Solvers, Correct Staters are predicted to be more likely than other Solvers to show abrupt improvement in two-alternative forced-choice performance (*Hypothesis 3*) and a decrease in response times (*Hypothesis 4*) after the last error.

3.3 Results

3.3.1 Questionnaire responses

Participants reported behaving in ways that have received little or no attention in the artificial-phonology-learning literature to date. To illustrate the contrast between what is often assumed to occur in a phonological-learning experiment and what our participants reported, we quote their own words before proceeding to a quantitative analysis.

Naïve participants, i.e., those who reported not having studied linguistics, were able to discover phonetic properties and invent ways to verbalise them, even for some properties which often take time and effort for Linguistics 101 students to grasp. Out of the 137 valid participants in this experiment, 36 (26%) did this. For example, the continuancy distinction (fricatives vs. stops) was intended by the experimenters to be non-verbalizable, but some participants recognised the feature and coined their own terminology:

> The feminine words used harsher consonant sounds and it was pretty clear from the beginning. Consonants p,d,t,etc were feminine whereas z,s,v, etc. sounds were masculine. (Participant fUlgjM, Explicit-Promoting, fricatives/stops)

> The words that ended more sharply seemed masculine than the feminine words. I followed the same rules as the first round here and looked for the same sounds. (Participant Explicit-Promoting, pzyaXQ, fricatives/stops)

The experimenters likewise intended place of articulation (labial vs. coronal) to be non-verbalizable, but one participant reported:

> The words had consonant sounds that were formed using the lips and front of the mouth. All of the studied words used "v," "p," "b," and "f" sounds, which are made with the lips and front of the mouth, so I chose the words that used those sounds (Participant XABNEW, labial/coronal)

Many participants verbalised a rule in the form of a list of letters, e.g.,

> I found that feminine words did not usually end in a t, z, or s. It usually ended with either an o or a u as the second to last letter, with usually an f or p as the last letter. (Participant PjMFZY, labial/coronal)

> I noticed that most of the words were pronounced starting with an o or a sound and often had a u sound somewhere in it. (Participant OUzBea, front/back)

> All words that I chose started with the "ah" sound. (Participant Mdantx, initial/second-syllable stress)

> Then, I noticed that when the second syllable was stressed I got the bell. (Participant SyzluI, initial/second-syllable stress)

Instead of three easily-verbalizable and three non-verbalizable features, as intended, the experiment turned out to have used one feature that was frequently verbalised as a feature (two vs. three syllables), two features that were frequently verbalised as letter lists (fricatives vs. stops and labials vs. coronals), one feature that was frequently verbalised ambiguously as a feature or a letter (initial vs. second-syllable stress; see Section 3.1.3), and two that were rarely verbalised (same vs. different consonants and front vs. back vowel). Summary statistics are shown in Table 5.

	Mentioned feature	Listed letters	Either
Intended verbalizable:			
Two vs. three syllables	0.59	0.00	0.59
Initial vs. second-syllable stress	—	—	0.62
All consonants identical vs. different	0.21	0.07	0.21
Intended non-verbalizable:			
Stressed vowel front vs. back	0.00	0.29	0.29
All consonants fricatives vs. stops	0.25	0.44	0.56
All consonants labial vs. coronal	0.08	0.62	0.62

Table 5: Empirical verbalizability of features in Experiment 1: proportion of all Rule-Seekers who mentioned the critical feature or listed letters in a correct or approximately-correct rule. (Every correct or approximately-correct rule either mentioned the feature, listed letters, or both; therefore, the "Either" column is also the proportion of Correct or Approximate Rule-Staters among the Rule-Seekers.) Report of stress location did not distinguish description from listing; see text.

Thus, despite experimenters' intentions, naïve participants may reason explicitly about phonetic properties, which they can discover during the experiment and for which they can invent phonetically non-arbitrary names to facilitate explicit reasoning. Additionally, even when the phonological stimuli are audio-only, as these were, participants may be mentally spelling them to facilitate explicit reasoning.

Nor do all participants report doing the experiment the same way (Table 6). Participants described a variety of approaches to the learning problem, and it often happened that an individual participant reported switching approaches during the experiment. Some examples follow.

	Intuiter		Non-Intuiter	
	Memoriser	Non-Memoriser	Memoriser	Non-Memoriser
Seeker	7	15	16	57
Non-Seeker	3	14	24	1

Table 6: Self-reported learning strategies (check-box responses), Experiment 1.

Pure intuition:

> I went by mostly similar sounds or letters used. No rules followed here just gut feeling. (Participant SaUkjT, Implicit-Promoting same/different consonants)

Pure sequential hypothesis testing:

> I considered different aspects of each word, such as number of syllables, the sounds of syllables, and what letters were used, and finally determined that for masculine words the last three letters were a consonant, a vowel, and the same consonant repeated, whereas with feminine words the last three letters were a consonant, a vowel, and then a different consonant. (Participant tIPXWj, Explicit-Promoting all consonants same/different)

Intuition and sequential hypothesis testing:

> I started mainly by intuition while trying to find patterns in apparent suffixes and prefixes. I also tried to find other patterns until I realized that the number of syllables appeared to denote the gender. I followed the pattern where two syllables equaled feminine and more than two equaled male. (Participant YnlqOd, Explicit-Promoting two/three syllables)

Intuition and rule of unknown origin:

> I tried vowel placement and sound but I don't know if thats how it works. So I went with my gut mostly. It seems the masculine is usually longer and sometimes with a long vowel in the middle with a lot of emphasis. (Participant RvWrHh, Explicit-Promoting two/three syllables condition)

Memorisation:

> I just tried to memorize the words by saying them out loud. Based on the words I was able to learn, I went off of those and chose words that sounded similar. (Participant DRrbim, Implicit-Promoting labial/coronal)

Tried rule-seeking but switched to memorisation:

> In the end, I just gave up and memorized which words were feminine and which weren't. I tried to find a pattern, for example, if words ended with a certain consonant, or if there were shorter or longer vowels and similar stuff, but honestly, there were no patterns I could discern. I didn't take any notes. I wasn't sure if you were allowed to. That might've been a good idea. I just tried to remember which words sounded feminine, even though I did not recognize a pattern. (Participant gbBIqh, Explicit-Promoting same/different consonants)

Focused attention on specific parts of the word:

> I first listened to the ending of the words to see if there was a pattern. Then, I noticed that when the second syllable was stressed I got the bell. The second syllable was stressed. (Participant SyzluI, Explicit-Promoting first/second syllable stress)

The reports differ from one participant to the next, even within a single condition, giving at least an initial impression that participants are sampled from a very mixed distribution. How seriously that impression is to be taken depends of course on how accurate self-report is, a question to which we now turn in the quantitative analysis. Self-report of cognitive processes is often viewed skeptically (Berry & Broadbent, 1984; Nisbett & Wilson, 1977), but it is often corroborated by objective behavioral measures, especially in intentional problem-solving tasks (Ericsson & Simon, 1980; Kellogg, 1982; Morris, 1981; White, 1988). One goal of this experiment series is to test the validity of self-report in phonological learning. The analysis, and the rest of this paper, will focus on rule-seeking and rule-stating, the bases of our hypotheses, rather than on memorisation.

3.3.2 Hypothesis 1: Rule-seeking and rule-stating are influenced (but not wholly determined) by instructions, feedback, and/or intention to learn

Results from all participants are plotted in Figure 3. Participants in the Explicit-Promoting condition were indeed significantly more likely than those in the Implicit-Promoting condition to be Rule-Seekers and Rule-Staters ($p = 0.0001643$ and 0.01053 respectively by Fisher's exact test, two-sided). A couple of Incorrect Staters performed well on the generalisation test, and so must have been basing their responses on something other than the incorrect rule they stated, perhaps intuition.[10]

3.3.3 Hypothesis 2: Stating a correct rule predicts better generalisation performance

Figure 3 also shows that participants tend to fall into two groups: Correct or Approximately-Correct Staters, who perform nearly perfectly on the generalisation test (black and gray circles), and Non-Staters or Incorrect Staters (empty and crossed circles), whose performance is widely distributed. In fact, *most* Correct or Approximately-Correct Staters (35/52) gave a pattern-conforming response on every single one of the 32 test trials, and *most* of those who gave 100% pattern-conforming responses (35/48) were Correct or Approximately-Correct Staters.

[10] As noted above, participants were assigned to the Stater category on the basis of their answer to the question "If you looked for a rule, what rules did you try?". Hence, the Incorrect Staters category also includes those who stated a rule that they said they tried and rejected.

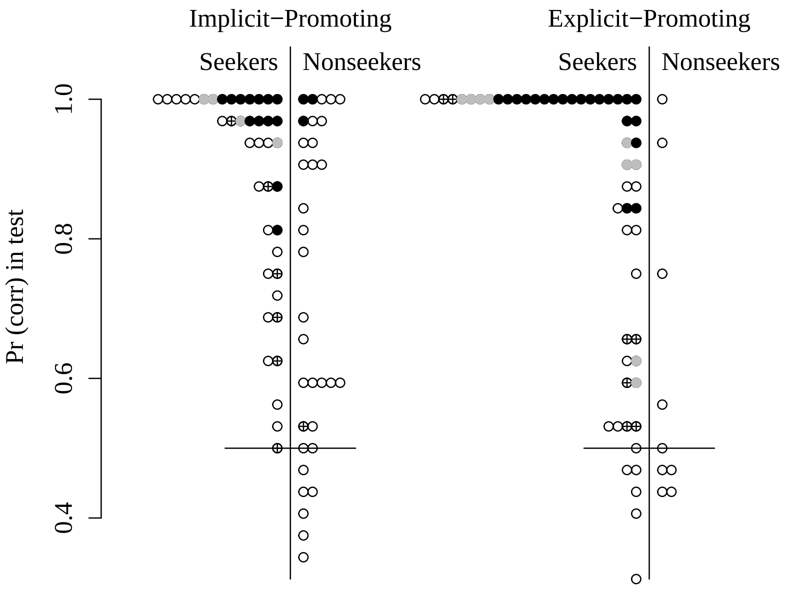


Figure 3: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 1. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

The effect of rule discovery on generalisation performance was quantified using *complex survey design* logistic regression with a two-stage sampling mode. This procedure, also known as a "population average model" or "sampler's model", treats each participant in the experiment as a cluster in a survey (e.g., a sample of size 100 voters in each U.S. State), and each 2AFC trial as a participant in the survey (e.g., an individual voter, responding to a single yes/no survey question). Complex survey design logistic regression is an alternative way of taking into account within-participant dependency (Bieler & Williams, 1995; Lumley & Scott, 2017; Williams, 2000) while avoiding convergence problems encountered when trying to fit mixed-effects logistic regression models to individual 2AFC responses. (The authors are indebted to Chris Wiesen of the Odum Institute for Social Science Research at the University of North Carolina, Chapel Hill, for suggesting this method.) Complex survey design logistic regression was used for all repeated-measures data in this article. The models were fit using the R package `survey` (Lumley, 2004,1; Lumley & Scott, 2017) with Training Group (0 = Explicit-Promoting, 1 = Implicit-Promoting), Rule Correctness (1 for Correct Staters, 0.5 for Approximate Staters, and 0 for others), and their interaction as fixed effects. The dependent variable was Correctness of each trial response (1 = pattern-conforming, 0 = non-conforming). The fitted model is shown in Table 7. The significant and positive intercept term means that even Incorrect Staters and Non-Staters performed above chance in the Explicit-Promoting condition, and the significantly positive coefficient for Implicit-Promoting means that they performed better in the Implicit-Promoting condition.

The large, highly significant coefficient for Rule Correctness, and the near-zero interaction term, mean that Correct and Approximate Staters did perform much better than Incorrect Staters and Non-Staters regardless of the training condition.

Coefficient	Estimate	Std. Error	t value	Pr(> $\|t\|$)	
(Intercept)	0.6552	0.1617	4.052	8.59e-05	***
Implicit-Promoting	0.4392	0.2185	2.010	0.0465	*
Rule Correctness	3.0614	0.4896	6.252	5.11e-09	***
Implicit-Promoting × Rule Correctness	−0.3707	0.7327	-0.506	0.6137	

Table 7: Summary of complex survey design logistic-regression model for pattern-conformity of generalisation-test responses, Experiment 1 (4384 responses from 137 participants).

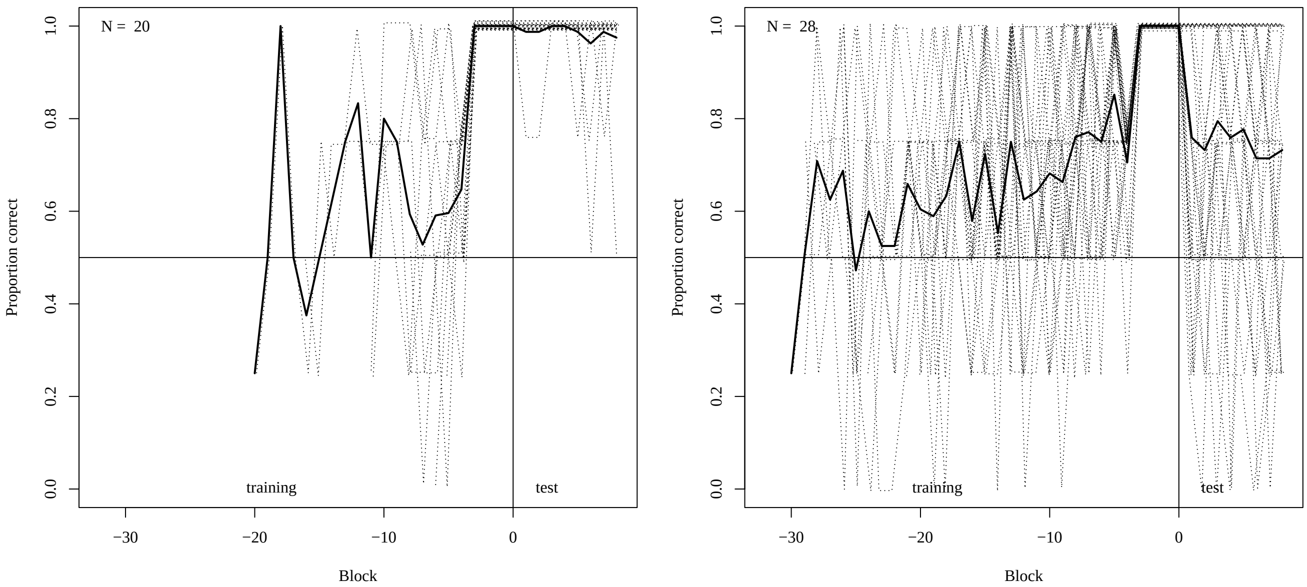
3.3.4 Hypotheses 3 and 4: Correct rule-stating is associated with more-abrupt learning curves and with response-time acceleration after the last error

The Explicit-Promoting condition yielded a learning curve for each participant, showing performance (proportion conforming responses) as a function of trial number. The curves for the Solvers (those who met the criterion of 4 consecutive correct 4-trial blocks before the end of the training phase) are shown in Figure 4. Performance in the 16-trial window preceding the last error was significantly lower for Correct and Approximate Staters than for other Solvers, as shown by the negative coefficient for *Rule Correctness* in the model of Table 8 (fitted using `svyglm`, as above, because of the repeated measure on Participants). This is as predicted by Hypothesis 3: Both the Correct Staters and the others learned the pattern to the same ultimate criterion level of 100%, but the transition was more abrupt (started from a lower baseline) for participants who stated a correct or partly-correct rule. Figure 4 also illustrates how near-perfect training performance in the test phase collapses when the participant does not state a correct rule (Hypothesis 2, above).

Hypothesis 4 was tested using trial-duration data from correct responses by Solvers in the Explicit-promoting condition. Only responses which occurred within sixteen trials before or after the last error were analyzed. Since response times on the very first trial of the experiment tended to be two or three times as long as on the second and subsequent trials, the very first trial was dropped if it occurred within the sixteen-

Coefficient	Estimate	Std. Error	z value	t value	Pr > $\|t\|$
(Intercept)	1.4514	0.1342	10.818	$1.39e-13$	***
Rule Correctness	−0.9662	0.2557	−3.779	0.000502	***

Table 8: Summary of the complex survey design logistic-regression model for pattern-conformity of training-phase responses in the 16-trial window preceding the last error before the 16-trial criterion run, for Solvers in the Explicit-Promoting condition of Experiment 1. (575 responses from 43 participants, excluding 5 more participants who either never made an error, or who only made an error on their first trial.)



Correct Staters

Others

Figure 4: Learning curves for Solvers in the Explicit-Promoting condition of Experiment 1, aligned to the last training error. (Since they are Solvers, the last training error precedes a 4-block sequence of correct responses.) Dashed lines are individuals, solid line is the mean across participants. Each point is the average of a block of four consecutive trials.

trial radius. Durations of less than 4 seconds or more than 30 seconds were excluded, which eliminated the most extreme 10% of responses. A general linear model was then fit using the same complex survey design used in other repeated-measures data in this paper via the R method `svyglm`, with log trial duration as the dependent variable. The critical predictors were *Preceding* (1 for trials preceding the last error, 0 for trials following it), *Rule Correctness* (1 for Correct Staters, 0.5 for Approximate Staters, else 0), and their interaction. Since Correct Staters' last error tended to occur earlier than other Staters', a nuisance variable, *log (trial number -1)*, was included to model out the overall shortening of response times after the (dropped) very first trial as the experiment progressed.[11]

Coefficient	Estimate	Std. Error	t value	Pr $> \|t\|$	
(Intercept)	2.43042	0.12625	19.251	$< 2e-16$	***
Preceding	-0.02132	0.02052	-1.039	0.30483	
Rule Correctness	0.03184	0.06427	0.495	0.62305	
Preceding × Rule Correctness	0.13217	0.05622	2.351	0.02375	*
log(Trial Number - 1)	-0.11067	0.02888	-3.831	0.00044	***

Table 9: Summary of the general linear model for log response time, correct responses from Solvers in the Explicit-Promoting condition within 16 trials of their last error. (1118 observations from 45 participants).

[11] Trial duration and trial number were natural-log-transformed to facilitate controlling for the acceleration of response times that is typically observed due to practice (Newell & Rosenbloom, 1981). The individual Solvers' log-trial-duration by log-trial-number plots were informally inspected to confirm that the transformation resulted in an approximately linear relation.

The fitted model is shown in Table 7. The intercept of about 2.5 and significant *Log trial number* coefficient mean that for Solvers who were not Correct or Approximate Staters, the time required to make a correct response shortened in a decelerating curve from about 12s on Trial 2 to a little less than 7s by Trial 128. The small, non-significant negative coefficient for *Preceding* means that for these participants, the 16 trials following the last error were not faster than those preceding it; if anything, they were a little slower, once the overall effect of *Log trial number* is corrected for. The small and nonsignificant effect of *Rule Correctness* means that when the other factors are controlled for, correctness of the stated rule had no significant effect on response time. Finally, the significant positive coefficient for the interaction between *Preceding* and *Rule Correctness* means that the more correct the stated rule was, the bigger the drop in response time between the trials preceding the last error and those following it. This is consistent with the effect described in non-linguistic learning by Haider & Rose (2007), in which rule discovery enables the participant to respond correctly after listening to only one of the two stimuli.

3.4 Discussion

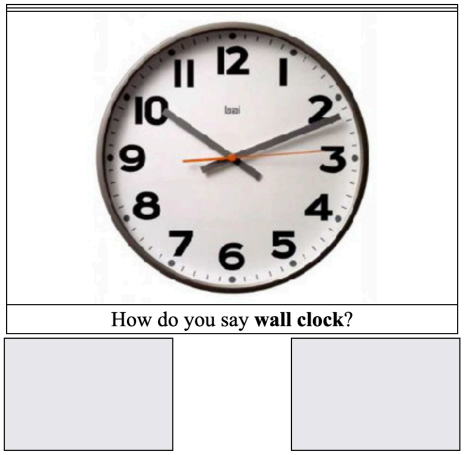
These results support the hypothesis that phonotactic patterns, like visual ones, can be induced using both implicit and explicit processes. The experiment also found learning-mode variety among participants. Although signs of explicit learning were rarer in the Implicit-Promoting condition than in the Explicit-Promoting condition, Rule-Seekers and Rule-Staters were found in substantial numbers in both conditions, and some participants reported using a mix of approaches. Many spontaneously used the alphabet or self-invented phonetic terminology to facilitate explicit learning.

4 Experiment 2

It is possible that Experiment 1 was not representative of phonological learning, either in the lab or in nature, and that it had characteristics that made both conditions especially favorable to explicit learning. Experiment 2 therefore differed from Experiment 1 in multiple ways. Where the gender-assignment scenario in Experiment 1 simulated learning to distinguish lexical classes within a language, Experiment 2 used a different scenario, vocabulary learning, to construct a situation in which participants could be asked to tell possible (well-formed) from impossible (ill-formed) words. Where the Implicit- and Explicit-Promoting conditions of Experiment 1 differed in instructions, feedback, and number of stimuli per trial, those of Experiment 2 differed only in whether each trial presented two well-formed stimuli (Implicit-Promoting) or one well-formed and one ill-formed (Explicit-Promoting). That made the feedback in the Implicit-Promoting condition of Experiment 2 useless for testing hypotheses about the pattern. One might therefore expect that the paradigm used in Experiment 2 would reduce or abolish the explicit learning observed in Experiment 1.

4.1 Methods

Participants in both conditions of Experiment 2 were trained to associate pictures with their (pattern-conforming) names, and were then shown novel pictures and asked to choose between novel pattern-conforming and nonconforming names for them. Instructions and feedback were the same in both training conditions. The only difference between the conditions was that the foil (incorrect choice) on each training trial was pattern-conforming in the Implicit-Promoting condition, and pattern-nonconforming in the Explicit-Promoting condition. A training trial is shown in Figure 5.



How do you say **wall clock**?

Figure 5: Participant view of a training and/or test trial in Experiment 2.

The critical feature was chosen from two/three syllables, first-/second-syllable stress, and stops/fricatives, featues which had all yielded high test-phase performance in Experiment 1. The training-phase instructions said nothing to either group about a pattern; participants where simply asked to learn which word went with which picture. Both training conditions in Experiment 2 used two-alternative choice trials with feedback. On each training trial, a positive word-picture pair (a picture plus a pattern-conforming word stimulus) was matched with a negative word-picture pair (a different picture plus a non-conforming word stimulus). The participant saw only the positive picture, with two buttons below it. Mousing over one button played the name of the picture (the positive stimulus); mousing over the other played a foil (the negative stimulus). After all 32 positive and all 32 negative pairs had been presented, the positive word-picture pairs were randomly re-matched with negative word-picture pairs for the next cycle (thereby changing, on average, all but one matching, Zager & Verghese 2007). The only difference between the training conditions was that the foils were pattern-conforming in the Implicit-Promoting condition, but non-conforming in the Explicit-Promoting condition.

The test phase for both groups was like the training phase for the Explicit-Promoting group, except that no feedback was given. Both groups were instructed to make their test-phase decision “based on which choice sounds more like it would be a word in the artificial language”. The Implicit-Promoting condition thus resembled other “artificial language” paradigms in which participants are familiarised on pattern-conforming items, then asked to choose between novel conforming and nonconforming items (e.g., Carpenter 2005; Chong 2021; Cristiá, Mielke, Daland & Peperkamp 2013; Finley 2011; Greenwood 2016; Kuo 2009; Lai 2015; Linzen & Gallagher 2014; Moreton 2008; Moreton, Pater & Pertsova 2015; Myers & Padgett 2014). Questionnaires were scored as in Experiment 1.

Of 229 participants who completed the experiment, 53 were excluded from analysis (4 reported a non-English L1, 5 reported taking written notes, 27 reported choosing test-phase responses that were maximally *un*like what they were trained on, 1 fell below the minimum performance criterion of at least 10 correct answers in the test phase, and 16 were excluded for two or more of these reasons), leaving 176 valid participants, 99 in the Explicit-Promoting condition and 77 in the Implicit-Promoting condition.[12]

4.2 Results

4.2.1 Hypothesis 1: Rule-seeking and rule-stating are influenced (but not wholly determined) by instructions, feedback, and/or intention to learn

Rule-Seekers and Rule-Staters were again found in both training conditions (Figure 6). Participants in the Explicit-Promoting condition were numerically more likely than those in the Implicit-Promoting conditon to be Rule-Seekers, but the difference was only marginally significant ($p = 0.08193$ by Fisher’s exact test, two-sided). Participants in the Explicit-Promoting condition were again significantly more likely to be Rule-Staters ($p = 0.0006625$ respectively by Fisher’s exact test, two-sided).

4.2.2 Hypothesis 2: Stating a correct rule predicts better generalisation performance

The data was analyzed using complex survey design, as in Experiment 1. Table 10 shows that Incorrect Staters and Non-Staters performed above chance in the Implicit-Promoting condition. Correct and Approximate Staters did much better than Incorrect Staters and Non-Staters in the Explicit-Promoting condition,

[12] In the interests of both statistical power and expositional brevity, the analysis here combines data from two temporally-separated runs of the same identical experiment, an initial batch with 142 completed participants (109 valid), plus a subsequent batch of 87 completed participants (67 valid) that was run alongside Experiment 5 to verify that the participant population was behaving stably on Type I. Results from both batches are very similar. The most notable consequence of merging them is that the facilitating effect of the Explicit-Promoting condition on rule seeking, which was significant when the initial batch is analyzed alone, drops to marginal significance when the two batches are analyzed together.

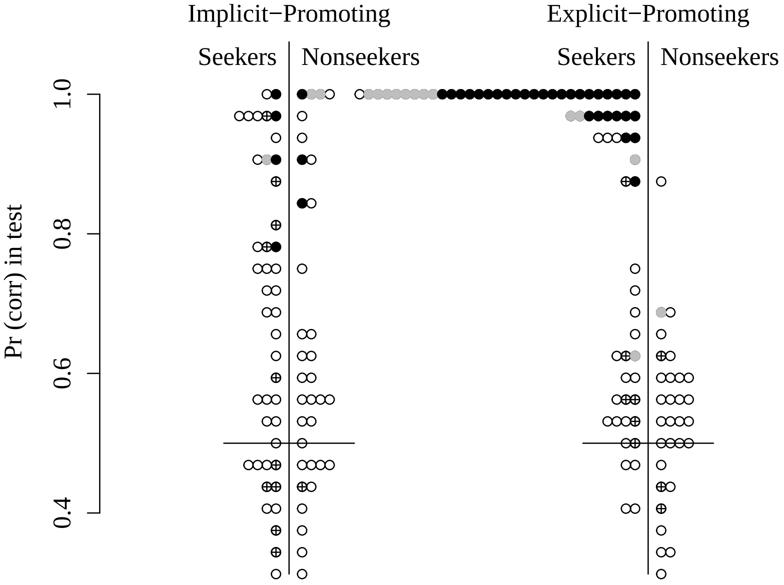


Figure 6: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 2. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

but the benefit of rule correctness vanished in the Implicit-Promoting condition, as shown by the significant negative coefficient for *Implicit-Promoting* × *Rule Correctness*.

Coefficient	Estimate	Std. Error	t value	*p*	
(Intercept)	0.137	0.085	1.615	0.1082	
Implicit-Promoting	0.456	0.181	2.519	0.0127	*
Rule Correctness	1.583	0.207	7.625	1.58e-12	***
Implicit-Promoting × Rule Correctness	-1.416	0.300	-4.715	4.97e-06	***

Table 10: Summary of fixed effects in the complex survey design logistic-regression model for pattern-conformity of generalisation-test responses, Experiment 2 (5632 responses from 176 participants).

4.2.3 Hypotheses 3 and 4: Correct rule-stating is associated with more-abrupt learning curves and with response-time acceleration after the last error

In the Explicit-Promoting condition, where attending to pattern-conformity could help performance, Correct Stater Solvers showed a more-abrupt performance jump across the last error, and their good performance persisted throughout the test phase. Others (Non-Staters, Incorrect Staters, and Approximate Staters) showed more-gradual improvement which tended to relapse in the test phase. The effect of Correct Stating on abruptness is confirmed statistically using the same model as in Experiment 1 (Table 11). The response-time acceleration at the last error as a function of Rule Correctness was replicated here (Table 12). Complex survey design was used as in Experiment 1 because of the repeated measure on Participant.

Coefficient	Estimate	Std. Error	t value	Pr > $\|t\|$	
(Intercept)	1.505	0.1916	7.858	6.98e-11	***
Rule Correctness	-0.731	0.2587	-2.827	0.00632	**

Table 11: Summary of the complex survey design logistic-regression model for pattern-conformity of training-phase responses in the 16-trial window preceding the last error before the 16-trial criterion run, for Solvers in the Explicit-Promoting condition of Experiment 2. (815 responses from 64 participants, excluding 3 more participants who either never made an error, or who only made an error on their first trial.)

Coefficient	Estimate	Std. Error	t value	Pr > $\|t\|$	
(Intercept)	2.539	0.1182	21.472	$< 2e - 16$	***
Preceding	-0.015	0.0240	-0.631	0.530642	
Rule Correctness	-0.175	0.0670	-2.623	0.010987	*
Preceding × Rule Correctness	0.106	0.0433	2.459	0.016795	*
log(Trial Number - 1)	-0.092	0.0241	-3.837	0.000298	***

Table 12: Summary of the general linear model for log response time, correct responses from Solvers in the Explicit-Promoting condition within 16 trials of their last error. (1646 observations from 66 participants.)

4.3 Discussion

The vocabulary-learning scenario of Experiment 2 produced nearly the same results as the gender-learning scenario of Experiment 1. The change in learning scenario thus did not affect the availability of implicit and explicit processes.

Unlike in Experiment 1, however, the two training conditions did not differ significantly in the rate of rule-seeking (perhaps because the instructions, task, and feedback were the same in both), and Correct and Approximate Stating, which in Experiment 1 occurred frequently in both training conditions, was in Experiment 2 confined almost entirely to the Explicit-Promoting condition. It thus appears that the opportunity to compare conforming and non-conforming stimuli on the same trial tends to facilitate successful explicit learning (not altogether surprisingly, since explicit learning relies on working memory; see Section 2).

5 Discussion: Experiments 1 and 2

The first two experiments asked whether human inductive learning of phonotactic patterns showed evidence for distinct implicit and explicit systems similar to that observed in inductive learning of non-linguistic patterns (Section 2).

	Explicit-Promoting		Implicit-Promoting	
	Seekers	Non-Seekers	Seekers	Non-Seekers
Non-Staters	17	9	18	29
Staters	37	0	28	4
Correct Staters	21	0	13	3
Approximate Staters	9	0	4	0
Incorrect Staters	7	0	6	1

Table 13: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 1

Hypothesis 1: Rule-seeking and rule-stating are influenced (but not wholly determined) by instructions, feedback, and/or intention to learn. This hypothesis was supported by differences between the Explicit- and Implicit-Promoting conditions in Experiments 1 and 2. However, in both experiments, Rule-Seekers were in the majority even in the Implicit-Promoting conditions, which were designed to discourage rule-seeking in the first place, to misdirect rule-seeking away from the actual pattern if attempted, and to render rule-seeking futile even if correctly directed. It may seem incredible that those who reported rule-seeking in either of the Implicit-Promoting conditions could have been doing anything that would benefit their performance in the generalisation test. And yet they were: Even in the Implicit-Promoting conditions, Rule-Seekers were significantly more likely than Non-Seekers to be Staters and to be Correct Staters (Tables 13 and 14; the four tables with the associated statistical tests, which were done using Firth-penalised logistic regression in order to reduce the risk of inflated significance due to empty or near-empty cells, are omitted for reasons of space).

	Explicit-Promoting		Implicit-Promoting	
	Seekers	Non-Seekers	Seekers	Non-Seekers
Non-Staters	20	26	28	28
Staters	49	4	15	6
Correct Staters	31	0	4	3
Approximate Staters	12	1	1	2
Incorrect Staters	6	3	10	1

Table 14: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 2

We can only speculate as to why there were so many Rule-Seekers in the Implicit-Promoting conditions, and what they were doing that could possibly have improved test-phase performance. Some may have only begun to look for a rule once the test phase started, but (mis)reported rule-seeking during training. However, it seems likely to us that many were simply doing what many of us would have been doing in their place, namely, trying to figure out what the experiment was really about. Even though the task made the search unhelpful for the training phase, participants may have noticed shared properties of the stimuli which they then used as the basis for a rule once the test phase started and they were confronted with non-conforming foils.

Hypothesis 2: Stating a correct rule predicts better generalisation performance. In both experiments, Correct and Partly-Correct Staters gave significantly more pattern-conforming responses on the generalisation test than did Non-Staters or Incorrect Staters. The effect was particularly clear among Solvers. All Solvers, by definition, finished the Explicit-Promoting training phase with sixteen consecutive correct responses, but the Correct Stater Solvers' high performance continued into the generalisation test, while that of the other Solvers fell sharply (see Figure 4).[13] Participants' rule reports were therefore largely accurate descriptions of their own response behavior. The straightforward interpretation is that participants responded by applying their stated rule.

Hypothesis 3: Correct rule-stating is associated with a more-abrupt learning curve. Solvers in the Explicit-Promoting condition had significantly lower performance immediately before their last error when they stated a correct rule than when they did not.

Hypothesis 4: Correct rule-stating is associated with response-time acceleration after the last error. This hypothesis was borne out. A straightforward interpretation of the two positive results is that rule discovery did have a shortening effect on response times, similar to that found in for non-linguistic learning by Haider & Rose (2007). Since two audio stimuli were presented on each training trial, one positive and one negative, rule discovery could have allowed a participant to respond after listening to only one of them.

6 Experiment 3

The implicit and explicit systems are hypothesised to have different architectures and hence different inductive biases: The rule-based explicit system is faster for patterns which depend on fewer features, while the cue-based implicit system is faster for patterns which are supported by multiple overlapping cues (see above, Section 2). Empirical support for this view comes from studies of visual pattern-learning involving the contrast between Shepard *et al.* (1961)'s "Type II" and "Type IV" patterns. A Type II pattern is an if-and-only-if relationship between two features, e.g., "circle if and only if black". A Type IV pattern is defined by resemblance to a three-feature prototype, e.g., "at most one feature different from a small white triangle" (Figure 7). The typical finding is that Type II patterns are easier for humans to learn inductively than Type IV (Nosofsky, Gluck, Palmeri, McKinley & Gauthier, 1994a; Shepard *et al.*, 1961; Smith *et al.*, 2004; Vigo, 2013).[14] Changing the experimental conditions so as to encourage implicit learning reduces performance on

[13] A mixed-effects logistic-regression model was fit to just the data from Solvers, with Correct Response (1/0) as the dependent variable, Correct Stater (1/0) as the predictor, and a random intercept for Participant. The coefficient for Correct Stater was 1.1 in Experiment 1 and 0.98 in Experiment 2, $p < 0.0001$ in both cases.

[14] *Linear separability* does not explain the difference: The same experiments find that Type II is also easier than the three-feature non-linearly-separable Type III; see also Medin & Schwanenflugel (1981); Moreton *et al.* (2017).

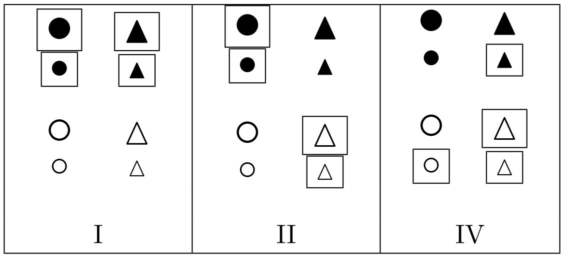


Figure 7: Examples of visual Type II and Type IV patterns.

Type II relative to Type IV (Kurtz *et al.*, 2013; Love, 2002; Minda *et al.*, 2008; Nosofsky & Palmeri, 1996; Rabi & Minda, 2016; Zettersten & Lupyan, 2020).

Several proposals have been advanced in the psychology literature to explain the observed advantage of Type II over Type IV. They are based on the idea that explicit rule learning is biased towards hypotheses that involve fewer relevant features. As noted in Section 2, the proposals differ as to how this bias comes about, a point which will become relevant in the post-hoc discussion (Section 9); for the nonce, we hypothesise merely that, since only two features are relevant for Type II, whereas three are relevant for Type IV, Type II has an advantage in explicit learning (Bradmetz & Mathy, 2008; Feldman, 2000,0; Kurtz *et al.*, 2013; Lafond, Lacouture & Mineau, 2007; Mathy & Bradmetz, 2004; Nosofsky *et al.*, 1994b; Shepard *et al.*, 1961; Vigo, 2009).

The two-systems hypothesis thus predicts that explicit learners will show an advantage for Type II over Type IV which will be reduced or reversed for implicit learners. If phonotactic learning uses the same two systems, the same effect of implicit versus explicit learning on the relative difficulty of Type II vs. Type IV ought to be observed. Some indication that this might be the case comes from studies by Gerken *et al.* (2019); Moreton *et al.* (2017), which found better performance on Type IV than Type II in adult phonotactic learning, perhaps (we conjecture) because the participants were learning implicitly. Those experiments did not, however, distinguish implicit from explicit learners, so we take up that task now.

Experiment 3 is like Experiment 1 except that, instead of all patterns being Type I (a single-feature affirmation), each participant receives either a Type II or a Type IV pattern. The two-systems theory predicts that participants who report explicit learning (rule-seeking) ought to show relatively better performance on Type II than Type IV as compared to participants who do not report explicit learning (*Hypothesis 5*).

6.1 Methods

The critical feature were chosen from among two/three syllables, stops/fricatives, and labials/alveolars. For each participant in the Type II condition, every phonological feature was randomly paired with a unique dimension of the Type II example in Figure 7; e.g., for one participant, stops/fricatives would be paired with black/white; for another, stops/fricatives might be paired with white/black, or with large/small, or with square/triangle. For each participant in the Type IV condition, the same was done with the Type IV example. Of 112 participants who completed the experiment, 31 were excluded from analysis (4 reported a non-English L1, 11 reported taking written notes, 7 reported choosing test-phase responses that were maximally *un*like what they were trained on, none fell below the minimum performance criterion of at least 10 correct answers in the test phase, and 2 were excluded for two or more of these reasons), leaving 88 valid participants, 19 in the Type II Implicit-Promoting condition, 16 in the Type IV Implicit-Promoting condition, 25 in the Type II Explicit-Promoting condition, and 28 in the Type IV Explicit-Promoting condition.

6.2 Results

Since no significant results were found in the analyses of Hypothesis 3 and Hypothesis 4 in this or in any subsequent experiment, the corresponding sections are omitted.

6.2.1 Hypothesis 1: Rule-seeking and rule-stating are influenced (but not wholly determined) by instructions, feedback, and/or intention to learn

Rule-seeking and rule-stating occurred in both training conditions (Figure 8). Fisher's exact test could no longer be used as it was in Experiments 1 and 2, because the additional Type (II vs. IV) factor meant that the data no longer formed a two-dimensional contingency table. In order to reduce the risk of false positives that arises when ordinary logistic regression is applied to data sets which have some cells with few observations in them, Firth-penalised logistic regression was used instead, fit using the `logistf` method in R's `logistf` package (Firth, 1993; Heinze & Ploner, 2018). *Seeker* was the dependent variable and *Training Condition* and *Type* were predictors. No significant effect of either predictor was found and no interaction (Table 15). However, in the Type II condition, Rule-Staters were significantly rarer in the Implicit-Promoting group as shown by the significant negative coefficient of *Implicit-Promoting* in Table 16 . No significant effects of or interactions with Type were found (Table 16).

Coefficient	Estimate	Std. Error	χ^2 value	p
(Intercept)	1.3156	0.4897	9.2122	0.00240
Implicit-Promoting	-0.3462	0.7097	0.2488	0.61790
IV	-0.2625	0.6527	0.1690	0.68099
IV × Implicit-Promoting	-0.9432	0.9713	0.9927	0.31906

Table 15: Fitted Firth-penalised logistic-regression model for Rule-Seeking as a function of Training Condition and Type, Experiment 3

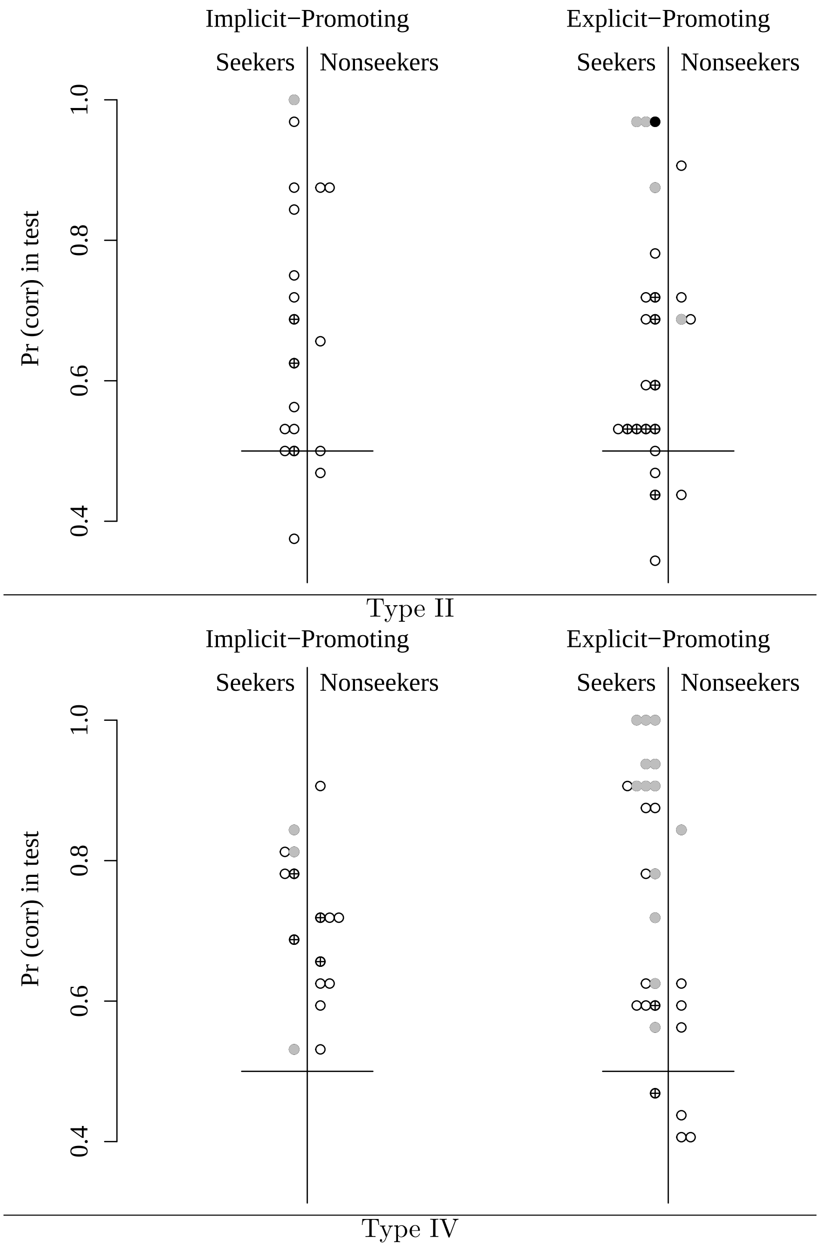


Figure 8: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 3. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

6.2.2 Hypothesis 2: Stating a correct rule predicts better generalisation performance

There were so few Correct and Approximate Staters in the Implicit-Promoting condition, particularly in Type II, that a model with *Rule Correctness* as a predictor could not be fit. The analysis was therefore restricted to the Explicit-Promoting condition alone. Pattern type was coded with Type II as 0 and Type IV as 1. Complex survey design logistic regression was used as in Experiment 1. The fitted model is shown

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	0.0769	0.4002	0.0384	0.8445	
Implicit-Promoting	-1.3137	0.6797	4.2464	0.0393	*
IV	0.0611	0.5511	0.0127	0.9099	
IV × Implicit-Promoting	0.9391	0.9268	1.0935	0.2956	

Table 16: Fitted Firth-penalised logistic-regression model for Rule-Stating as a function of Training Condition and Type, Experiment 3

in Table 17. Participants in the Type II condition who were not Correct or Approximate Staters nonetheless chose pattern-conforming responses at above-chance levels, as shown by the significantly positive intercept. Those who were Correct or Approximate Staters were very much more likely to respond in conformity with the pattern, as shown by the large and significant positive coefficient for *Rule Correctness*. Participants in Type IV did not differ significantly from those in Type II.

Coefficient	Estimate	Std. Error	z value	Pr($>|z|$)	
(Intercept)	0.3928	0.1256	3.128	0.0029	**
Rule Correctness	3.0925	0.8564	3.611	0.0007	***
IV	0.1092	0.2189	0.499	0.6202	
Rule Correctness × IV	−0.5352	1.1195	−0.478	0.6347	

Table 17: Summary of fitted complex survey design logistic-regression model for pattern-conformity of generalisation-test responses, Experiment 3, Explicit-Promoting condition only. Type II is the reference category. (1696 responses from 53 participants.)

6.2.3 Hypothesis 5: Rule-seeking is associated with better IFF/XOR and/or worse Family-resemblance performance

Previous experiments with non-linguistic patterns have found that performance on Type II patterns is typically better than on Type IV, and that conditions which favor explicit learning improve performance on Type II relative to Type IV (Kurtz *et al.*, 2013; Love, 2002). Figure 8 shows that in both the Explicit-Promoting and Implicit-Promoting groups, Seekers perform better than Non-Seekers on Type IV, but not on Type II. I.e., Type II, the pattern type that in the past has been found to benefit the most from an explicit learning approach, actually benefited the least. Among Seekers in both training conditions, performance on Type II is well below that on Type IV. These observations are confirmed by a mixed-effects logistic-regression model (Table 18), in which the only significant terms are the intercept and the interaction *IV* × *Seeker*. The hypothesis is therefore, not merely not supported, but outright contradicted by the results.

Coefficient	Estimate	Std. Error	z value	$\Pr(>\lvert z\rvert)$	
(Intercept)	0.78846	0.31239	2.524	0.01358	*
IV	-0.57335	0.38415	-1.493	0.13950	
Implicit-Promoting	-0.05757	0.47608	-0.121	0.90405	
Seeker	-0.17628	0.35987	-0.490	0.62559	
IV × Implicit-Promoting	0.58286	0.54828	1.063	0.29095	
IV × Seeker	1.28717	0.47689	2.699	0.00848	**
Seeker × Implicit-Promoting	0.18239	0.55661	0.328	0.74401	
Seeker × Implicit-Promoting × IV	-0.93507	0.68775	-1.360	0.17777	

Table 18: Summary of fixed effects in the fitted logistic-regression model for pattern-conformity of generalisation-test responses, Experiment 3. Type II is the reference category. (2816 responses from 88 participants.)

6.3 Discussion

The learning-mode variety found with Type I patterns in Experiments 1 and 2 was replicated here: Rule-seeking and rule-stating occurred in both training conditions and for both Type II and Type IV target patterns, and the Explicit-Promoting condition facilitated rule-stating. Moreover, learning mode affected, not just the absolute, but the *relative* difficulty of the two pattern types: Self-reported rule-seeking improved performance on Type IV so much that it exceeded performance on Type II. Learning mode is thus confirmed to vary between participants and to affect inductive bias. The direction of the effect (explicit learning favoring Type IV) contradicts the prediction of Hypothesis 5, being unexpected under models of rule-based learning which incorporate a bias towards patterns that depend on fewer features (Section 2). A post-hoc explanation for this surprising reversal is deferred to Section 9 below.

The Correct Staters, who in the earlier experiments formed a mode at 100% in the distribution of test-phase performance, were absent from Experiment 3, presumably because the correct rules were harder to find or to state. The Approximately-Correct Staters did show better generalisation performance than Non-Staters and Incorrect Staters, as before. However, no significant effect of Rule Correctness on abruptness or response time was found. That could simply be because Rule Correctness only ranged up to 0.5, i.e, any helpful effect of Rule Correctness was coming from a less-helpful partially-correct rule. More interestingly, it could instead be a sign that multi-feature rules are found incrementally rather than all at once: If rule discovery occurs in successive stages (e.g., with the identification of one relevant feature at a time), then each stage would bring with it a increment in accuracy and a decrement in response time, so that any comparison of performance just before and just after a single trial would find only a small difference. The lack of a Rule Correctness effect on abruptness or response time could also mean that multi-feature rules, once found, are harder to apply, such that the difference in accuracy or speed between having no rule at all and having a correct rule is smaller when the correct rule is hard to apply than when it is easy to apply.

7 Experiment 4

The results of Experiment 3 were surprising enough that Experiment 4 was done to see if they would replicate. In Experiment 3, some of the Type II patterns, and all of the Type IV patterns, involved the two features fricatives/stops and labial/coronal, which were both realised on the consonants. That meant that some Type II patterns could be learned correctly by focusing on the consonants and learning only the consonant inventory, whereas no Type IV pattern could be learned correctly without integrating features that were spread across the stimulus. To remove this asymmetry between Type II and Type IV, Experiment 4 used first- vs. second-syllable stress in place of Experiment 3's labial vs. coronal consonants. Otherwise the two experiments were the same.

7.1 Methods

The stimuli, instructions, and procedure were identical to those of Experiment 3. Of 173 participants who completed the experiment, 4 were subsequently excluded for reporting a non-English L1, 1 for reporting deliberately choosing test-phase items that sounded different from the training items, 7 for reporting taking written notes, and 2 for falling below the 10-out-of-32 criterion. That left 151 valid participants, 36 in the Explicit-Promoting Type II condition, 40 in the Implicit-Promoting Type IV condition, 40 in the Explicit-Promoting Type II condition, and 35 in the Explicit-Promoting Type IV condition.

7.2 Results

7.2.1 Hypothesis 1: Rule-seeking and rule-stating are influenced (but not wholly determined) by instructions, feedback, and/or intention to learn

As in all previous experiments, rule-seeking and rule-stating occurred in both training conditions and in both pattern-type conditions (Figure 9). A Firth-penalised logistic-regression model with *Seeker* as the dependent variable and *Training Condition* and *Type* as predictors was used, for the same reasons explained in Section 6.2.1, and found no significant effect of either predictor and no interaction (table omitted to save space). Implicit-Promoting Type II participants were numerically less likely than Explicit-Promoting Type II participants to state a rule, but the difference was only marginally significant (Table 19). Participants in the Type IV Explicit-Promoting condition were much more likely to be Staters than those in the Type II Explicit-Promoting condition, and those in the Type IV Implicit-Promoting condition did not differ significantly from those in the Type IV Explicit-Promoting condition.

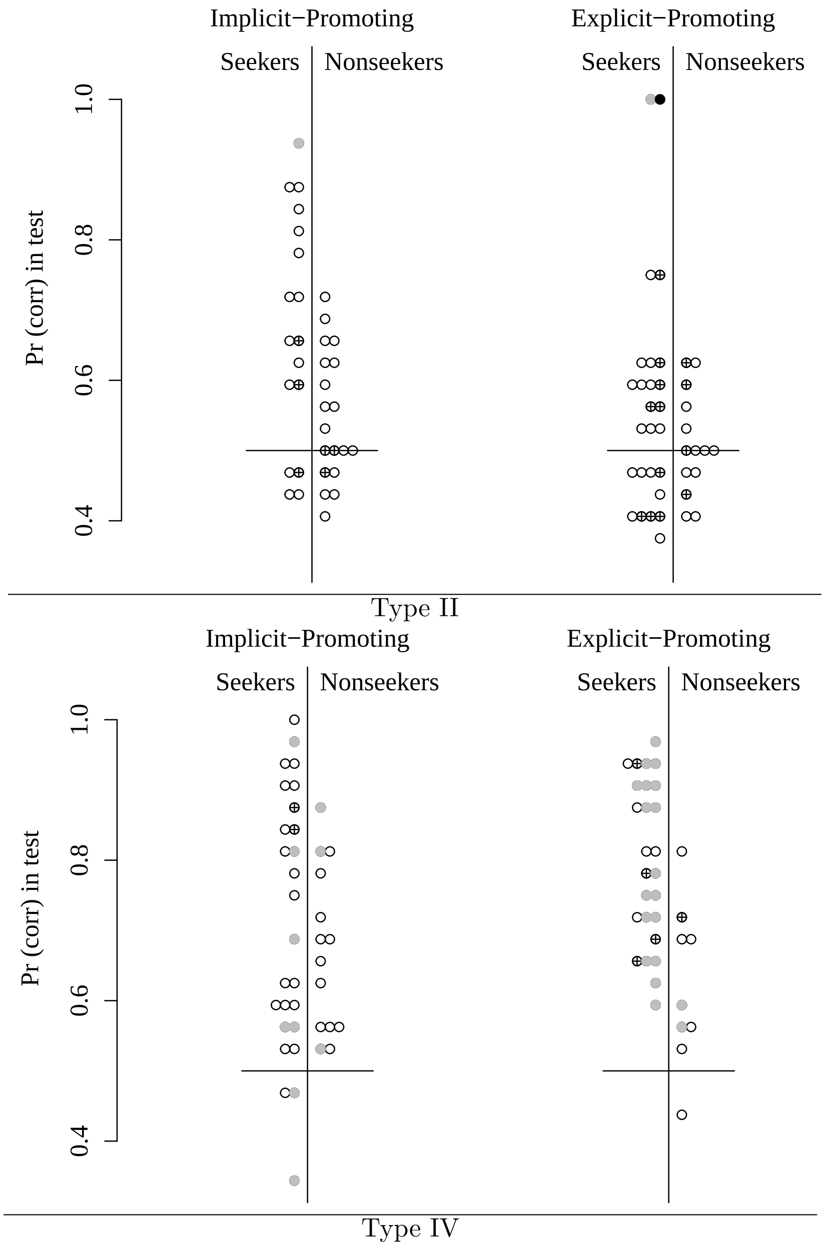


Figure 9: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 4. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	-0.4978	0.3260	2.4638	0.1164	
Implicit-Promoting	-0.8716	0.5274	2.9275	0.0870	.
IV	1.2541	0.4875	7.1490	0.0075	**
IV × Implicit-Promoting	-0.7088	0.7263	0.9728	0.3239	

Table 19: Fitted Firth-penalised logistic-regression model for Rule-Stating as a function of Training Condition and Type, Experiment 4.

7.2.2 Hypothesis 2: Stating a correct rule predicts better generalisation performance

The mode at 100% pattern-conforming generalisation responses which was found in Experiments 1 and 2, and which disappeared with the switch to more-complex pattern types in Experiment 3, was again absent here. There were not enough Correct or Approximate Staters in the Type II condition for the model to be fit accurately, so only the Type IV condition was analyzed. A complex survey design logistic-regression model was used because of the repeated measure on Participant. Neither Rule Correctness nor Training Condition had any significant influence on test-phase performance (table omitted to save space). The ineffectiveness of Rule Correctness may be due in part to its small range: Since there were no Correct Staters, the experiment could only measure the (smaller) difference between Approximately-Correct Staters and others.

7.2.3 Hypothesis 5: Rule-seeking is associated with better IFF/XOR and/or worse Family-resemblance performance

The fitted model is shown in Table 20. In the Explicit-Promoting condition, Seekers do not outperform Non-Seekers in the Type II condition (as shown by the small and non-significant coefficient for *Seeker*), but do so in the Type IV condition (large and significant coefficient for *IV* × *Seeker*). This much is consistent with what was found in Experiment 3. In the Implicit-Promoting condition, however, this interaction is significantly reduced (the large and significantly nonzero coefficient for the three-way interaction is numerically larger than the coefficient for *IV* × *Seeker*).

7.3 Discussion

The outcome of Experiment 4 was very much like that of Experiment 3. In particular, the same novel effect seen in Experiment 3 is replicated in Experiment 4: In the Explicit-Promoting condition, self-reported rule-seeking benefits Type IV performance *more* than it does Type II performance, contrary to previous theoretical proposals and unlike previous experimental results. This is true even though no Seekers in the Type IV condition succeeded in stating a wholly correct rule, and even though Approximate Stating did not significantly improve generalisation performance. These results again directly contradict Hypothesis 5.

Coefficient	Estimate	Std. Error	t value	Pr($>|t|$)	
(Intercept)	0.03572	0.07436	0.480	0.63172	
IV	0.46032	0.17022	2.704	0.00768	**
Implicit-Promoting	0.16229	0.11175	1.452	0.14862	
Seeker	0.24005	0.14738	1.629	0.10556	
IV × Implicit-Promoting	0.05834	0.23292	0.250	0.80256	
IV × Seeker	0.64570	0.25332	2.549	0.01186	*
Seeker × Implicit-Promoting	0.29953	0.24421	1.227	0.22201	
Seeker × Implicit-Promoting × IV	-0.98736	0.38759	-2.547	0.01191	*

Table 20: Summary of fitted complex survey design logistic-regression model for pattern-conformity of generalisation-test responses, Experiment 4. Type II is the reference category. (4832 responses from 151 participants.)

8 Experiment 5

This experiment sought to replicate the rule-seeking effect on the Type IV advantage over Type II using the vocabulary-learning paradigm of Experiment 2.

8.1 Methods

The stimuli, instructions, and procedure were identical to those of Experiment 2, except that each participant was randomly assigned a Type II, or Type IV pattern, stated in terms of two or three of the properties disyllabic/trisyllabic, first-/second-syllable stress, and stop/fricative consonants. 176 participants completed the experiment. 8 were subsequently excluded for reporting a non-English L1, 31 for reporting deliberately choosing test-phase items that sounded different from the training items, 7 for reporting taking written notes, 3 for falling below the 10-out-of-32 criterion, and 8 for multiple reasons. That left 119 valid participants, 33 in the Implicit-Promoting Type II condition, 31 in the Implicit-Promoting Type IV condition, 22 in the Explicit-Promoting Type II condition, and 33 in the Explicit-Promoting Type IV condition.

8.2 Results

8.2.1 Hypothesis 1: Rule-seeking and rule-stating are influenced (but not wholly determined) by instructions, feedback, and/or intention to learn

As in all previous experiments, both rule-seeking and rule stating were found in both training conditions (Figure 10). However, as in Experiment 2, training condition did not affect either of these two variables significantly (tables omitted to save space).

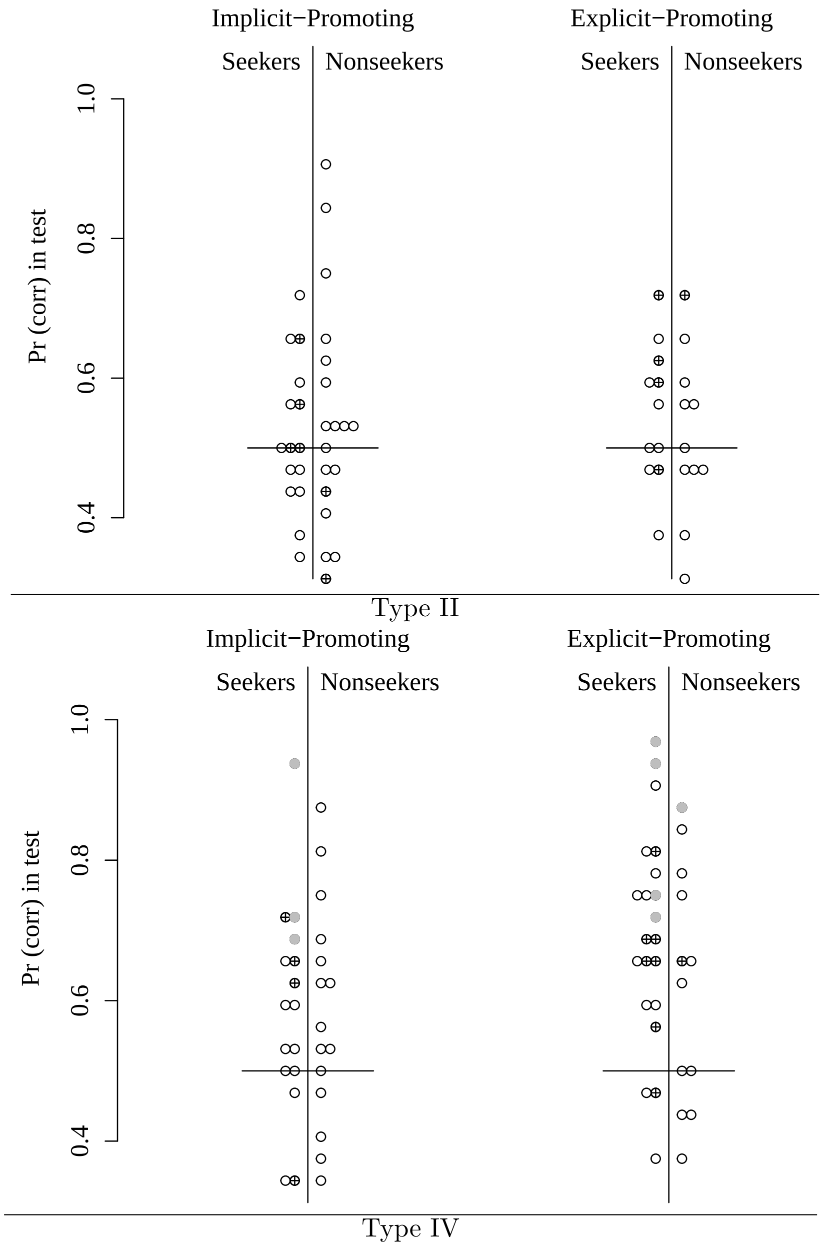


Figure 10: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 5. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

8.2.2 Hypothesis 2: Stating a correct rule predicts better generalisation performance

Because there were no Correct or Approximate Staters in the Type II condition, the effects of *Rule Correctness* on pattern-conformity of test-phase responses were analyzed only for Type IV. A complex survey design logistic-regression model with the pattern-conformity of each generalisation-test response as the dependent variable and *Rule Correctness* and *Training Condition* was fit as shown in Table 21. The large and highly-significant coefficient for *Rule Correctness* shows that Correct and Approximate Stating increased the chances of a pattern-conforming test-phase response in the Explicit-Promoting condition. Incorrect Staters and Non-Staters in the Explicit-Promoting condition were marginally less likely to give a pattern-conforming response, but the coefficient for the interaction between *Training Condition* and *Rule Correctness* was small and non-significant, indicating that Correct and Approximate Stating facilitated pattern-conforming test-phase responses in both training conditions.

Coefficient	Estimate	Std. Error	t value	Pr($>|t|$)	
(Intercept)	0.5539	0.1187	4.666	1.76e-05	
Rule Correctness	2.3614	0.7432	3.177	0.00235	***
Implicit-Promoting	−0.2935	0.1583	−1.854	0.06863	.
Rule Correctness × Implicit-Promoting	−0.3362	1.0816	−0.311	0.75700	

Table 21: Summary of fitted logistic-regression model for pattern-conformity of generalisation-test responses, Experiment 5, Type IV only. (2048 responses from 64 participants.)

8.2.3 Hypothesis 5: Rule-seeking is associated with better IFF/XOR and/or worse Family-resemblance performance

Rule-seeking had no significant effect on test-phase pattern-conformity of Type II vs. Type IV in either training condition. There was a non-significant numerical trend in the same direction as in Experiments 3 and 4 (table omitted to save space).

8.3 Discussion

The results of Experiment 5 did not directly contradict Hypothesis 5 the way Experiments 3 and 4 did, but Experiment 5 did not support Hypothesis 5 at all. The nonsignificant numerical trend went in the "wrong" direction for Hypothesis 5, i.e., to the benefit of Seekers over Non-Seekers in the Type IV condition but not in the Type II condition.

9 Discussion: Experiments 3–5

Participants in the Type II/IV experiments (Experiments 3–5), like those in the Type I experiments (Experiments 1 and 2), showed evidence of using both implicit and explicit learning. Rule-seeking and rule-stating were found in every condition of every experiment, and were facilitated in the Explicit-Promoting condition in Experiments 3 and 4 relative to the Implicit-Promoting condition. Some findings of the Type I experiments were not replicated. Correct Staters were much rarer; the generalisation test no longer showed a mode at or near 100% corresponding to Correct and Approximate Staters; and Correct or Approximate Stating no longer resulted in significantly more-abrupt learning curves or faster response times among Solvers. These differences from the Type I experiments can be traced to the same source: Since the completely correct rule is harder to find and state explicitly in the Type II/IV experiments than in the Type I experiments, any effect of rule correctness in the Type II/IV experiments originates mainly in the weaker effect of an approximately-correct rule.

The results also confirmed the prediction that implicit and explicit learning can have different inductive biases. What was surprising was the direction of the difference: Rule-seeking benefited performance in the Type IV condition, but not the Type II condition (in Experiment 3 and the Explicit-Promoting condition of Experiment 4). In these experiments, the explicit and implicit processes are roughly equally successful on Type II (no significant effect of *Seeker* or interaction with it), but the implicit process is less successful than the explicit process on Type IV (significant positive interaction of *Seeker* with *Type = IV*). The effect of learning mode on bias was the exact opposite of what one would expect based on the theories and empirical studies of domain-general explicit learning reviewed in Section 2.

Where does this difference between phonological learning and non-linguistic learning come from? We consider two possible post-hoc hypotheses, one based on between-domain differences in implicit learning, the other based on between-domain differences in explicit learning.

9.1 Option 1: Implicit phonological learning is feature-minimizing

One possibility is that *implicit* learning works differently in phonology versus other domains. Specifically, a human learner might have a domain-general explicit learning process with a feature-minimisation bias, a domain-general implicit learning process without a feature-minimisation bias, and a dedicated implicit phonological learning process with an especially strong feature-minimisation bias that pre-empts the general implicit process for phonological stimuli. Feature-minimisation bias is a well-established idea in phonology, e.g., Chomsky & Halle (1968, 168, 221, 331, 334); Bach & Harms (1972); King (1969, 88–89); Smith (1973, 155–158); Gordon (2004); Hayes (1999); Hayes & White (2013); Hayes & Wilson (2008); Hayes, Zuraw, Siptár

& Londe (2009); Kiparsky (1982); Pycha, Nowak, Shin & Shosted (2003), to name only some proposals that explicitly ascribe a bias to human learners of natural language. [15] That would explain how switching from implicit to explicit learning can improve performance on Type IV relative to Type II in phonology, while doing the opposite in other domains.

However, there is empirical evidence against this alternative: In an in-person study that compared eight-feature phonological patterns with their feature-by-feature visual analogues using a task similar to the Implicit-Promoting condition of Experiment 1, performance was significantly better on Type IV than on Type II in both the phonological and the visual condition (Moreton *et al.*, 2017, Experiments 1 and 2). Since Option 1 hypothesizes that both the phonology-specific implicit process and the domain-general explicit process learn Type II better than Type IV, there is no way for Option 1 to explain this outcome, regardless of how many participants used each learning mode.

9.2 Option 2: Explicit learning is impeded by irrelevant features

A second possibility is that *explicit* learning works differently in phonology versus other domains — not because humans are endowed with a dedicated phonology-specific explicit learning mechanism, but because phonological stimuli have properties which are rarely found in other domains and which interact with the explicit process of serial hypothesis testing.

To see how this might happen, we note that the explicit-learning component of the domain-general two-systems hypothesis (Section 2) is based on data mainly from experiments in low-dimensional stimulus spaces like that in Figure 7, where the only features that vary are colour, shape, and size. Phonological stimuli are different. The ones used here varied not only on the six experimentally-manipulated dimensions of syllable count, labial vs. coronal, etc., but also on features like voicing and vowel height that were randomized to make distinct stimuli. They also varied on ad-hoc phonological properties such as "ends with an F", which participants readily invented. Typical phonological stimuli thus have many more pattern-irrelevant features than the non-linguistic patterns which formed the empirical basis of the explicit-learning component of the two-systems theory. Irrelevant features have been shown to increase errors and time- or trials-to-criterion in non-linguistic concept-learning tasks that are designed to encourage explicit reasoning (e.g., Archer, Bourne & Brown 1955; Keele & Archer 1967; Kepros & Bourne 1966; Peterson 1962). Here we propose a way in which they may also influence sensitivity to Type II vs. Type IV.

[15] Others propose a negative correlation between feature count and human learnability in the lab (e.g., Durvasula & Liter 2020) or between feature count and typological frequency in natural language (e.g., Clements 1985; Halle 1961; Sagey 1990, 1; Kenstowicz 1994, 21; Clements 2003), but do not say outright that human learners are subject to a bias in acquiring natural language. The preceding examples do not include the many feature-minimisation proposals that are intended, not as a description of learning biases or typological asymmetries, but as an analytic criterion for the guidance of linguists. See Chen (1973) for discussion of the distinction.

Suppose that explicit learners search for the relevant dimensions ("attribute identification", Haygood & Bourne 1965) by serially testing one-dimensional rules (Neisser & Weene, 1962; Wattenmaker, McQuaid & Schwertz, 1995). In the Type IV condition, this is a promising strategy: Each relevant dimension, individually, can yield a one-feature rule that is 75% correct during Explicit-Promoting training, and that characterizes 75% of the (all-positive) training items during Implicit-Promoting training. In contrast, one-feature rules based on the irrelevant dimensions are only 50% correct. A learner in the Type IV condition can use this difference in correctness rate to distinguish relevant from irrelevant dimensions. But in the Type II condition, any single relevant dimension yields a rule that is only 50% correct, thus making the relevant dimensions indistinguishable from the irrelevant ones. The serial-search procedure is bound to fail. One Type II participant described the failure thus:

> I looked for many different kinds of rules to no avail. I tried going by the vowel at the beginning of the word. I tried going by what consonants were used, how many syllables, what consonants were used when certain numbers of syllables were used, the long and short sounds of vowels, and anything else I could think of. I couldn't find a rule. From then on I decided to go more for gut feeling and finally I began to focus on memorizing the words. (Participant AJvCRg, Experiment 3, Type II, Explicit-Promoting)

Support for this explanation comes from the fact that in all three II/IV experiments, Seekers in the Type IV condition were much more likely than those in the Type II condition to mention at least one of the pattern-relevant features in their free-response answers (Table 22). Across all three experiments, there was a grand total of 2 Correct and 8 Approximate Type II Staters out of 103 Seekers, versus 0 Correct and 53 Approximate Type IV Staters out of 117 Seekers (Types II and IV significantly different from each other by Fisher's exact test, odds ratio = 0.13, $p < 10^{-8}$). Seekers in the Type IV condition, it seems, readily identified at least one relevant feature, whereas those in the Type II condition could hardly find the relevant features at all.

	Experiment		
Type	3	4	5
II	0.03	0.07	0.00
IV	0.36	0.40	0.14

Table 22: Proportion of valid Seekers mentioning at least one pattern-relevant feature in any free-response question.

The results of Experiments 3– 5 can then be interpreted as follows: The implicit parallel system, being non-voluntary, is used by all participants, and learns Type IV better than Type II. The explicit serial system, when impeded by irrelevant features, rarely succeeds on Type II, but often finds a one-feature approximation to Type IV, giving Type IV a further boost among participants who voluntarily use the explicit system.[16]

10 General discussion

The principal conclusions of the present study are that phonotactic learning, like non-linguistic learning, can happen implicitly or explicitly, and that the implicit and explicit processes can have different inductive biases (Table 23). These conclusions are strengthened when we note that the experiments could not cleanly sort participants into one group of exclusively explicit learners and another of exclusively implicit ones: What these experiments detected, they detected by comparing a less-explicit group with a more-implicit group. These conclusions converge with and extend other recent findings on the existence of implicit vs. explicit processes in phonological learning (Chen 2021; Kimper 2016; Moreton & Pertsova 2016; Moreton *et al.* 2021).

One surprising finding was the unexpected direction of the effect of rule-seeking on inductive bias: More-explicit learning facilitated learning of Type IV patterns relative to Type II instead of hindering it. As discussed in Section 9, that result cannot be explained by positing a dedicated phonology-specific implicit learning process; instead, it has implications for the explicit component of the two-systems theory. Any adequate theory of domain-general explicit learning will have to explain explicit phonological learning as well, and so will need to take into account the effects of irrelevant features.

	Experiment				
	1	2	3	4	5
Hypothesis	G	V	G	G	V
1: Task affects rule-seeking/-stating,	✓	✓[a]	?/✓[b]	—	—
but staters and non-staters found in both tasks (% staters, EP:IP):	59:36	59:27	36:20	52:43	52:51
2: Correct rule stating facilitates generalisation performance	✓	✓	?/✓[c]	?/—[d]	✓
3: Correct rule stating associated with more-abrupt learning curve	✓	✓	—	—	—
4: Correct rule stating associated with RT acceleration after last error	✓	✓	—	—	—
5: Rule-seeking facilitates Type II relative to Type IV	n/a	n/a	→←	→←	—

Table 23: Summary of hypotheses tested in Experiments 1–5. G = gender scenario, V = vocabulary scenario. ✓ = supported, — = not supported, →← = contradicted (significant result in *opposite* direction). EP = Explicit-Promoting condition. IP = Implicit-Promoting condition.

[a] One effect only marginally significant.

[b] Found for rule-stating for Type II only.

[c] Found only for Explicit-Promoting condition; not enough rule staters in Implicit-Promoting.

[d] Not enough rule staters in Type II, and no Correct Staters to analyze in Type IV.

[16] Additional evidence that some participants were using the implicit system simultaneously with the explicit one comes from the fact that, although Type IV participants' rule statements often mentioned only one pattern-relevant feature, their responses were not based solely on a one-feature rule. A one- (or two-) feature rule, consistently applied, would have produced 75% correct test-phase performance. It is clear from Figures 8, 9, and 10, that there is no mode at 75% in the distribution of proportion correct for Seekers in the Type IV condition. If anything, there tends to be a notch near 75%, and the mode among the Approximate Staters (gray dots) is well above 85%. Type IV Staters' responses must therefore be based on something more than just their explicitly-stated approximate rule.

10.1 What goes on in phonological learning experiments?

Because learning experiments have come to play a major role in testing phonological theories, it is important to understand just what is happening in them. Researchers' assumptions may be wrong in consequential ways. For example, the authors were surprised by participants' ability to verbalise phonological features (Section 3.3.1), and reviewers were surprised by participants' successful rule-seeking in conditions that were designed to discourage, misdirect, and frustrate it (Section 5). At a more mundane, but no less consequential, level, participants, especially in Internet-based experiments, may be using other non-psychological resources that experimenters would not know about without asking; e.g., 5.1% of our participants reported taking written notes. (Their data was excluded from the analysis, as noted above in the individual "Results" sections, but that was only possible because the questionnaire specifically asked about note-taking.)

To be sure, it may be that all of the paradigms used in this study were abnormally favorable to explicit learning, whereas those used in other studies elicited only implicit learning of precisely the sort that is responsible for natural first- or second-language acquisition. That would be very fortunate. However, we will not know for certain until we have a better understanding of what participants in phonological-learning experiments are actually doing. It is therefore important to scrutinise experimental paradigms for signs of learning-mode variety. The methods used in this paper are one attempt at doing that, and other proposed methods for distinguishing implicit from explicit learning can be found in the literatures on non-linguistic learning and second-language learning that go far beyond simply asking whether participants can state the correct rule (see references in Section 2, above, as well as, e.g., Rebuschat 2013; Tunney & Shanks 2003). What we find by applying them may affect the interpretation of previous studies, and may give researchers better control over future ones.

10.2 Ecological validity of lab-learned phonology

Most of the evidence bearing on the ecological validity of phonological learning in the lab (i.e., does it use the same processes as natural L1 or L2 learning?) comes from comparing inductive biases in the lab with typological asymmetries in natural language. The linking hypothesis motivating these comparisons is that the more closely lab-learning biases agree with natural-language typological asymmetries, the more likely it is that both reflect the influence of Universal Grammar. The agreement between typology and lab results is not strikingly close.

In terms of abstract pattern structure, learners in the lab tend to favor Type I over Type IV and Type IV over Type II (Gerken *et al.* 2019; Moreton & Pater 2012a,1; see also Glewwe 2019, 168f.). In natural language, phonologically-active classes defined by fewer features are indeed more common (Mielke, 2004,0)

— but on the other hand, phonologically-active classes that can be expressed as Type II, like the left panel in Table 24, are more common than those that can be expressed as Type IV, like the right panel (Moreton & Pertsova, 2014).

	−voice		+voice	
	−distr	+distr	−distr	+distr
−cont	t	p, tʃ, k	$\boxed{\text{n}}$	$\boxed{\text{m}}$
+cont	$\boxed{\text{s}}$	$\boxed{\text{ʃ, x, h}}$	l	w,j

Type II: Consonants of Unami Delaware (Goddard, 1979). Boxes enclose sounds that can precede non-coronal stops; they are [+cont] iff [−voice].

	−round		+round	
	−back	[+back]	−back]	+back
+high	$\boxed{\text{i}}$	$\boxed{\text{ɨ}}$	$\boxed{\text{y}}$	u
−high	$\boxed{\text{e}}$	a	ø	o

Type IV: The boxes enclose those Kirghiz vowels which undergo raising and tensing before palatal consonants (Hebert & Poppe, 1963, 3–7). The set can be described as "any sound within one feature of /i/".

Table 24: Examples of natural-language Type II and Type IV patterns, found by analyzing P-Base (Mielke, 2008; Moreton & Pertsova, 2014). Features as in Chomsky & Halle (1968).

In terms of phonetic substance, phonetically-motivated patterns are the norm in natural language, and it is the phonetically "unnatural" patterns that linguists regard as demanding special explanation (Anderson, 1981; Bach & Harms, 1972; Brohan & Mielke, 2018; Buckley, 2000). In the lab, however, substantive biases — those discriminating between patterns on the basis of phonetic motivation, such as final obstruent devoicing vs. final obstruent voicing — are weak relative to structural biases (Moreton & Pater 2012a; Moreton & Pater 2012b). (For opposing views, see Chen 2020; Finley 2017; Hayes & White 2013; Lin 2023; Martin & Peperkamp 2020.) If substantive biases exist at all, they may be restricted to particular experimental conditions such situations of high uncertainty (Baer-Henney, Kügler & van de Vijver, 2015; Huang & Do, 2022) or perceptual unclarity as in casual speech (Greenwood, 2016), or to particular kinds of phonetic motivation such as perception rather than production (Glewwe, 2019), or they may only emerge when the phonetic motivation is especially strong (Glewwe, 2022).

One interpretation of these mismatches between lab biases and typology, both in terms of abstract structure and in terms of phonetic substance, is that typical short-term phonological-learning experiments are ecologically invalid; i.e., the learning processes they are "about" are not the same ones used by natural L1 or L2 learners. If that is so, then making the experiments more lifelike ought to change the outcomes in a direction that more closely matches what is observed in natural-language typology. One example of this line of work is Martin & Peperkamp (2020), which compared the learning of a typologically common vs. a typologically rare phonological pattern with vs. without sleep between training and testing (more vs. less lifelike). That study found no difference between the two conditions, but it could simply be that more is needed for adequate ecological validity than a night's sleep — perhaps even as much as is needed to acquire phonotactic patterns in a natural second language (e.g., Trapman & Kager 2009).

An alternate interpretation of the lab-vs.-typology mismatches is that it is the linking hypothesis that is wrong, and that some interfering factor prevents natural-language typology from being an accurate reflection of biases in natural-language learning. A likely candidate for that other factor is asymmetries in the phonetic precursors available for phonologisation (Blevins, 2004; Hyman, 1976; Ohala, 1993). Suppose that inductive bias favors Type IV over Type II patterns, such that, given two phonetic precursors, one a continuous analogue of Type II, the other of Type IV, the probability of phonologising the Type II pattern from the same level and duration of exposure is 0.01, while the like probability for the Type IV pattern is 0.05. If Type II precursors outnumber Type IV precursors by 20 to 1 across the languages of the world, then phonologisation will create four times as many new Type II patterns as Type IV, despite the fivefold inductive bias in the opposite direction. That hypothesis could be tested by looking for discrepancies between what is available for phonologisation and what actually gets phonologised. That could be done across languages, by comparing phonetic typology with phonological typology to see if certain phonological patterns are systematically underrepresented in relation to their phonetic precursors (Cole & Iskarous, 2001; Hombert, Ohala & Ewan, 1979; Moreton, 2008; Myers & Padgett, 2014), or across time, by comparing known precursors with their phonologised forms to see if phonologisation is unfaithful to precursors in systematic ways (Hayes 1999).

10.3 Phonotactic learning as concept learning

The present results give no reason to think that implicit or explicit phonotactic learning in the laboratory is anything but a special case of domain-general concept learning, using domain-general processes which only appear to be unique because the phonological stimulus space has properties rarely studied elsewhere, such as many pattern-irrelevant features (Section 9) or multiple instances of the same feature within a single stimulus Moreton 2012, 167–168). The conclusion of Section 9, that learners in the present experiments seemed to be serially testing candidate features for relevance, thus supports models of concept learning in which features are serially tested for relevance, such as the mental-model theory (Goodwin & Johnson-Laird, 2011,1), RULEX (Nosofsky *et al.*, 1994b), or, within phonology, the proposal of Durvasula & Liter (2020, 210), over models in which single- and many-feature candidate rules are tested simultaneously, such as Rational Rules (Goodman *et al.*, 2008).

However, the domain-specificity or otherwise of learning in phonology is a very large question that will not be settled by a handful of experiments. We therefore urge researchers to investigate more parallels or differences between phonological and non-phonological learning. Clear and convincing evidence of substantive bias would argue for a special status for phonological learning, but there are many other avenues to explore. Do the biases seen in the acquisition and use of non-linguistic patterns show up in analogous phonological-learning experiments? Do they affect natural first- or second-language acquisition, or leave their imprint on natural-language typology? Within language, is phonological learning biased in a different way from morphological, lexical, or syntactic learning? These questions can only be answered by thoroughgoing comparative study of inductive biases in analogous problems across domains (Moreton *et al.*, 2017).

Acknowledgments

The authors are indebted to Rachel Broad, Will Carter, Caleb Hicks, and Josh Fennell for assistance with stimulus manufacture and questionnaire scoring; to Kenneth Kurtz, Joe Pater, Lisa Sanders, Jen Smith, and the UNC-Chapel Hill P-Side Caucus for discussion and comments; to Chris Wiesen of the Odum Institute for Social Science Research at UNC-Chapel Hill for statistical advice; to the anonymous reviewers and *Phonology* editors for their helpful scrutiny and suggestions; and to audiences at the Linguistics Association of Great Britain (2014), the Deutsche Gesellschaft für Sprachforschung (2015), the Boston University Conference on Language Development (2015), the Manchester Phonology Meeting (2015), UC Santa Cruz (2016), MIT (2016), UCLA (2017), the Linguistic Society of America (2018), and Rutgers University (2018). The research was funded in part by a grant to the authors from the U.S. National Science Foundation, Grant # BCS 1651105.

References

Anderson, John R. (1991). The adaptive nature of human categorization. *Psychological Review* **98**. 409–429.

Anderson, Nathaniel D. & Gary S. Dell (2018). The role of consolidation in learning context-dependent phonotactic patterns in speech and digital sequence production. *Proceedings of the National Academy of Sciences* **115**. 3617–3622.

Anderson, Stephen R. (1981). Why phonology isn't "natural". *Linguistic Inquiry* **12**. 493–539.

Archer, E. James, Lyle E. Bourne & Frederick G. Brown (1955). Concept identification as a function of irrelevant information and instructions. *Journal of Experimental Psychology* **49**. 153–164.

Ashby, F. Gregory, Leola A. Alfonso-Reese, And U. Turken & Elliott M. Waldron (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review* **105**. 442–481.

Ashby, F. Gregory, Erick J. Paul & W. Todd Maddox (2011). COVIS. In Emmanuel M. Pothos & Andy J. Willis (eds.) *Formal approaches in categorization*, chapter 4. Cambridge, England: Cambridge University Press, 65–87.

Bach, Emmon & Robert T. Harms (1972). How do languages get crazy rules? In R. P. Stockwell & R. K. S. Macaulay (eds.) *Linguistic change and generative theory*, chapter 1. Bloomington: Indiana University Press, 1–21.

Baer-Henney, Dinah, Frank Kügler & Ruben van de Vijver (2015). The interaction of language-specific and universal factors during the acquisition of morphophonemic alternations with exceptions. *Cognitive Science* **39**. 1537–1569.

Batterink, Laura J., Ken A Paller & Paul J. Reber (2019). Understanding the neural bases of implicit and statistical learning. *Topics in Cognitive Science* **11**. 482–503.

Batterink, Laura J., Paul J. Reber, Helen J. Neville & Ken A. Paller (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language* **82**. 62–78.

Berry, Diane C. & Donald E. Broadbent (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology* **36A**. 209–231.

Bieler, Gayle S. & Rick L. Williams (1995). Cluster sampling techniques in quantal response teratology and developmental toxicity studies. *Biometrics* **51**. 754–776.

Blevins, Juliette (2004). *Evolutionary phonology.* Cambridge: Cambridge University Press.

Bley-Vrooman, R. (1990). The logical problem of foreign language learning. *Linguistic Analysis* **20**. 3–49.

Boersma, Paul & David Weenink (2013). PRAAT Version 5.3.14. Software, www.praat.org.

Boersma, Paul & David Weenink (2021). PRAAT Version 5.3.14. Software, www.praat.org.

Bower, Gordon H. & Thomas R. Trabasso (1964). Concept identification. In Richard C. Atkinson (ed.)

Studies in mathematical psychology, chapter 2. Stanford, California: Stanford University Press, 32–74.

Bradmetz, Joël & Fabien Mathy (2008). Response times seen as decompression times in boolean concept use. *Psychological Research* **72**. 211–234.

Breen, Gavan & Rob Pensalfini (1999). Arrernde: a language with no syllable onsets. *Linguistic Inquiry* **30**. 1–25.

Brohan, Andrew & Jeff Mielke (2018). Frequent segmental alternations in P-Base 3. In Larry M. Hyman & Frans Plank (eds.) *Phonological typology.* Berlin and Boston: Walter de Gruyter, 196–228.

Brown, Janessa L. (2009). *A brief sketch of Urama grammar with special consideration of particles marking agency, aspect, and modality.* Master's thesis, University of North Dakota, Grand Forks.

Brown, Jason, Alex Muir, Kimberly Craig & Karika Anea (2016). *A short grammar of Urama.* Number 32 in Asia-Pacific Linguistics. Canberra, Australia: Australian National University.

Bruner, Jerome S., Jacqueline J. Goodnow & George A. Austin (1956). *A study of thinking.* New York: John Wiley and Sons.

Buckley, Eugene (2000). On the naturalness of unnatural rules. In *Proceedings from the Second Workshop on American Indigenous Languages*, volume 9 of *UCSB Working Papers in Linguistics*. 1–14.

Carpenter, Angela C. (2005). Acquisition of a natural vs. an unnatural stress system. In Allejna Brugos, Manuella R. Clark-Cotton & Seungwan Han (eds.) *Papers from the 29th boston university conference on language development (bucld 29).* Somerville: Cascadilla Press, 134–143.

Carpenter, Angela C. (2006). *Acquisition of a natural versus an unnatural stress system.* PhD dissertation, University of Massachusetts, Amherst.

Carpenter, Angela C. (2010). A naturalness bias in learning stress. *Phonology* **27**. 345–392.

Carpenter, Angela C. (2016). Learning natural and unnatural phonological stress by 9- and 10-year-olds: a preliminary report. *Journal of Child Language Acquisition and Development* **4**. 62–77.

Chen, Matthew (1973). On the formal expression of natural rules in phonology. *Journal of Linguistics* **9**. 223–249.

Chen, Tsung-Ying (2020). An inductive learning bias toward phonetically driven tonal phonotactics. *Language Acquisition* **27**. 331–361.

Chen, Tsung-Ying (2021). On the learnability of level-based and unit-based OCP generalizations: an artificial grammar learning study. *Glossa: a journal of general linguistics* **7**. 1–45.

Chomsky, Noam & Morris A. Halle (1968). *The sound pattern of English.* Cambridge, Massachusetts: MIT Press.

Chong, Adam J. (2021). The effect of phonotactics on alternation learning. *Language* **97**. 213–244.

Ciborowski, Tom & Michael Cole (1972). A cross-cultural study of conjunctive and disjunctive concept learning. *Child Development* **43**. 774–789.

Ciborowski, Tom & Michael Cole (1973). A developmental and cross-cultural study of the influences of rule structure and problem composition on the learning of conceptual classifications. *Journal of Experimental Child Psychology* **15**. 193–215.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook* **2**. 225–252.

Clements, G. N. (2003). Feature economy in sound systems. *Phonology* **20**. 287–333.

Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational Psychology and Measurement* **20**. 37–46.

Cole, Jennifer S. & Khalil Iskarous (2001). Effects of vowel context on consonant place identification: implications for a theory of phonologization. In Elizabeth Hume & Keith Johnson (eds.) *The role of speech perception in phonology.* San Diego: Academic Press, 103–122.

Corbett, Greville G. (1991). *Gender.* Cambridge Textbooks in Linguistics. Cambridge, England: Cambridge University Press.

Cristiá, Alejandrina, Jeff Mielke, Robert Daland & Sharon Peperkamp (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology* **4**. 259–285.

DeKeyser, R. (2003). Implicit and explicit learning. In C.J. Doghty & M.H. Long (eds.) *The handbook of second language acquisition.* Oxford: Blackwell, 314–348.

Dell, Gary S.., David R. Adams & Antje S. Meyer (2000). Speech errors, phonotactic constraints, and implicit learning: a study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**. 1355–1367.

Dell, Gary S, Amanda C Kelley, Suyeon Hwang & Yuan Bian (2021). The adaptable speaker: a theory of implicit learning in language production. *Psychological review* **128**. 446–487.

Do, Youngah, Elizabeth Zsiga & Jonathan Havenhill (2016). Naturalness and frequency in implicit phonological learning. Slides from a talk at the 90th Annual Meeting of the Linguistic Society of America.

Ọmọruyi, Thomas O. (1986). Adjectives and adjectivalization processes in Ẹdo. *Studies in African Linguistics* **17**. 283–302.

Durvasula, Karthik & Adam Liter (2020). There is a simplicity bias when generalising from ambiguous data. *Phonology* **37**. 177–213.

Ellis, Nick C. (1994). *Implicit and explicit learning of languages.* London: Academic Press.

Ericsson, K. Anders & Herbert A. Simon (1980). Verbal reports as data. *Psychological Review* **87**. 215–251.

Evans, Jonathan St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* **49**. 255–278.

Feldman, Jacob (2000). Minimization of Boolean complexity in human concept learning. *Nature* **407**.

630–633.

Feldman, Jacob (2006). An algebra of human concept learning. *Journal of mathematical psychology* **50**. 339–368.

Finley, Sara (2011). Generalization to novel consonants in artificial grammar learning. In *Proceedings of the 33rd annual conference of the cognitive science society, ed. by laura carlson, christoph hoeschler and thomas f. shipley*. 318–23.

Finley, Sara (2017). Learning metathesis: evidence for syllable structure constraints. *Journal of Memory and Language* **92**. 142–157.

Finley, Sara & William Badecker (2010). Linguistic and non-linguistic influences on learning biases for vowel harmony. In S. Ohlsson & R. Catrambone (eds.) *Proceedings of the 32nd annual conference of the cognitive science sciety*. Austin, Texas: Cognitive Science Society, 706–711.

Firth, David (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**. 27–38.

Franco, Karlien, Eline Zenner & Dirk Speelman (2018). Let's agree to disagree: (variation in) the assignment of gender to nominal anglicisms in Dutch. *Journal of Germanic Linguistics* **30**. 43–87.

Gamer, Matthias, Jim Lemon, Ian Fellows & Puspendra Singh (2019). *Irr: various coefficients of interrater reliability and agreement*. R package version 0.84.1.

Gerken, LouAnn, Carolyn Quam & Lisa Goffman (2019). Adults fail to learn a type of linguistic pattern that is readily learned by infants. *Language Learning and Development* .

Glewwe, Eleanor (2019). *Bias in phonotactic learning: experimental study of phonotactic implicationals*. PhD dissertation, University of California, Los Angeles.

Glewwe, Eleanor (2022). Substantive bias and the positional extension of major place contrasts. *Glossa: a journal of general linguistics* **7**. 1–37.

Gluck, Mark A. & Gordon H. Bower (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General* **117**. 227–247.

Goddard, Ives (1979). *Delaware verbal morphology: a descriptive and comparative study*. Taylor and Francis.

Goodman, Noah, Joshua B. Tenenbaum, Jacob Feldman & Thomas L. Griffiths (2008). A rational analysis of rule-based concept learning. *Cognitive Science* **32**. 108–154.

Goodwin, G. P. & P. N. Johnson-Laird (2011). Mental models of Boolean concepts. *Cognitive Psychology* **63**.

Goodwin, Geoffrey P. & Philip N. Johnson-Laird (2013). The acquisition of Boolean concepts. *Trends in Cognitive Sciences* **17**. 128–133.

Gordon, Matthew (2004). Syllable weight. In Bruce Hayes, Robert Kirchner & Donca Steriade (eds.) *Phonetically-based phonology*. Cambridge, England: Cambridge University Press, 277–312.

Green, Peter S & Karlheinz Hecht (1992). Implicit and explicit grammar: an empirical study. *Applied Linguistics* **13**. 168–184.

Greenwood, Anna (2016). *An experimental investigation of phonetic naturalness.* PhD dissertation, University of California, Santa Cruz.

Greer, G. Brian (1979). A study of rule-learning in children using set operations. *Journal of Experimental Child Psychology* **28**. 174–189.

Haider, Hilde & Michael Rose (2007). How to investigate insight: a proposal. *Methods* **42**. 49–57.

Halle, Morris (1961). On the role of simplicity in linguistic description. In Roman Jakobson (ed.) *Structure of language and its mathematical aspects.* American Mathematical Society, 89–94.

Hayes, Bruce (1999). Phonetically driven phonology: the role of optimality in inductive grounding. In Michael Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer & Kathleen Wheatly (eds.) *Functionalism and formalism in linguistics*, volume 1: General Papers. Amsterdam: John Benjamins, 243–285.

Hayes, Bruce & James White (2013). Phonological naturalness and phonotactic learning. *Linguistic inquiry* **44**. 45–75.

Hayes, Bruce & Colin Wilson (2008). A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* **39**. 379–440.

Hayes, Bruce, Kie Zuraw, Péter Siptár & Zsuzsa Londe (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language* **85**. 822–863.

Haygood, R. C. & L. E. Bourne (1965). Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review* **72**. 175–195.

Hebert, Raymond J. & Nicholas Poppe (1963). *Kirghiz manual*, volume 33. Bloomington: Indiana University Press.

Heinze, Georg & Meinhard Ploner (2018). *Logistf: firth's bias-reduced logistic regression.* R package version 1.23.

Høiland-Jørgensen, Toke, Bengt Ahlgren, Per Hurtig & Anna Brunstrom (2016). Measuring latency variation in the Internet. In *Proceedings of the 12th international conference on emerging networking experiments and technologies.* 473–480.

Hombert, Jean-Marie, John J. Ohala & William G. Ewan (1979). Phonetic explanations for the development of tones. *Language* **55**. 37–58.

Huang, Tingyu & Youngah Do (2022). Substantive bias and variation in the acquisition of/n/˜/l/alternation. In *Proceedings of the annual meetings on phonology*, volume 9.

Hyman, Larry M. (1976). Phonologization. In Alphonse Juilland (ed.) *Linguistic studies offered to Joseph*

Greenberg: second volume: phonology. Saratoga, California: Anma Libri, 407–418.

Jakobson, Roman C., Gunnar M. Fant & Morris Halle (1952). *Preliminaries to speech analysis: the distinctive features and their correlates.* Cambridge, Massachusetts: MIT Press.

Keele, Steven W. & E. James Archer (1967). A comparison of two types of information in concept identification. *Journal of Verbal Learning and Verbal Behavior* **6**. 185–192.

Kellogg, Ronald T. (1982). When can we introspect accurately about mental processes? *Memory and Cognition* **10**. 141–144.

Kelly, John (1969). Vowel patterns in the Urhobo noun. *Journal of West African Languages* **6**. 21–26.

Kenstowicz, Michael (1994). *Phonology in generative grammar.* Cambridge, Massachusetts: Blackwell.

Kepros, Peter C. & Lyle E. Bourne (1966). Identification of biconditional concepts: effect of number of relevant and irrelevant dimensions. *Canadian Journal of Psychology/Revue Canadienne de Psychologie* **20**. 198–207.

Keren, Gideon & Yaacov Schul (2009). Two is not always better than one: a critical evaluation of two-systems theories. *Perspectives on psychological science* **4**. 533–550.

Kim, Robyn, Aaron Seitz, Heather Feenstra & Ladan Shams (2009). Testing assumptions of statistical learning: is it long-term and implicit? *Neuroscience Letters* **461**. 145–149.

Kimper, Wendell (2016). Asymmetric generalisation of harmony triggers. In Gunnar Hansson, Ashley Farris-Trimble, Kevin McMullin & Douglas Pulleyblank (eds.) *Proceedings of the 2015 Annual Meeting on Phonology.*

King, Robert D. (1969). *Historical linguistics and generative grammar.* Englewood Cliffs, New Jersey: Prentice-Hall.

Kiparsky, Paul (1982). Linguistic universals and linguistic change. In *Explanation in phonology.* Dordrecht: Foris, 13–43.

Krashen, Stephen (1982). *Principles and practice in second language acquisition.* Oxford Pergamon.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review* **99**. 22–44.

Kuo, Li-Jen (2009). The role of natural class features in the acquisition of phonotactic regularities. *Journal of psycholinguistic research* **38**. 129–150.

Kurtz, Kenneth J., Kimery R. Levering, Roger D. Stanton, Joshua Romero & Steven N. Morris (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition* **39**. 552–572.

Lafond, Daniel, Yves Lacouture & Guy Mineau (2007). Complexity minimization in rule-based category learning: revising the catalog of Boolean concepts and evidence for non-minimal rules. *Journal of*

Mathematical Psychology **51**. 57–75.

Lai, Regine (2015). Learnable vs. unlearnable harmony patterns. *Linguistic Inquiry* **46**. 425–451.

Lai, Yeeking Regine (2012). *Domain specificity in learning phonology.* PhD dissertation, University of Delaware.

Landis, J. Richard & Gary G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**. 159–174.

Lee, Yuh-Show (1995). Effects of learning contexts on implicit and explicit learning. *Memory and Cognition* **23**. 723–734.

Lewandowsky, Stephan (2011). Working memory capacity and categorization: individual differences and modelling. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **37**. 720–738.

Lichtman, Karen M. (2013). Developmental comparisons of implicit and explicit language learning. *Language Acquisition* **20**. 93–108.

Lichtman, Karen Melissa (2012). *Child-adult differences in implicit and explicit second language learning.* PhD dissertation, University of Illinois at Urbana-Champaign.

Lin, Yu-Leng (2023). Are human learners capable of learning arbitrary language structures. *Brain Sciences* **13**. 181.

Lindahl, Maj-Britt (1964). The importance of strategy in a complex learning task. *Scandinavian Journal of Psychology* **5**. 171–180.

Linzen, Tal & Gillian Gallagher (2014). The timecourse of generalization in phonotactic learning. In *Proceedings of the annual meetings on phonology*, volume 1.

Love, Bradley C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review* **9**. 829–835.

Love, Bradley C. & Arthur B. Markman (2003). The nonindependence of stimulus properties in human category learning. *Memory and Cognition* **31**. 790–799.

Love, Bradley C., Douglas L. Medin & Todd M. Gureckis (2004). SUSTAIN: a network model of category learning. *Psychological Review* **111**. 309–332.

Lumley, Thomas (2004). Analysis of complex survey samples. *Journal of Statistical Software* **9**. 1–19.

Lumley, Thomas (2019). `survey`: analysis of complex survey samples, R package version 3.35–1. Comprehensive R Archive Network, http://cran.r-project.org.

Lumley, Thomas & Alastair Scott (2017). Fitting regression models to survey data. *Statistical Science* **32**. 265–278.

Maddox, W. Todd & F. Gregory Ashby (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes* **66**. 309–332.

Maddox, W. Todd, J. Vincent Filoteo & J. Scott Lauritzen (2007). Within-category discontinuity interacts with verbal rule complexity in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **33**. 197–218.

Martin, Alexander & Sharon Peperkamp (2020). Phonetically natural rules benefit from a learning bias: a re-examination of vowel harmony and disharmony. *Phonology* **37**. 65–90.

Mathews, Robert C., Ray R. Buss, William B. Stanley, Fredda Blanchard-Fields, Jeung Ryeul Cho & Barry Druhan (1989). Role of implicit and explicit processes in learning from examples: a synergistic effect. *Journal of Experimental Psychology* **15**. 1083–1100.

Mathy, Fabien & Joel Bradmetz (2004). A theory of the graceful complexification of concepts and their learnability. *Current Psychology of Cognition/Cahiers de Psychologie Cognitive* **22**. 41–82.

McHugh, Mary L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica* **22**. 276–282.

Medin, Douglas L. & Paula J. Schwanenflugel (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory* **7**. 355–368.

Mielke, Jeff (2004). *The emergence of distinctive features.* PhD dissertation, Ohio State University.

Mielke, Jeff (2008). *The emergence of distinctive features.* Oxford, England: Oxford University Press.

Minda, John Paul, Amy S Desroches & Barbara A Church (2008). Learning rule-described and non-rule-described categories: a comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **34**. 1518–1533.

Moore-Cantwell, Claire, Joe Pater, Robert Staubs, Benjamin Zobel & Lisa Sanders (2017). Event-related potential evidence of abstract phonological learning in the laboratory. MS, Department of Linguistics, University of Massachusetts, Amherst. (Under review.).

Moreton, Elliott (2008). Analytic bias and phonological typology. *Phonology* **25**. 83–127.

Moreton, Elliott (2012). Inter- and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language* **67**. 165–183.

Moreton, Elliott & Joe Pater (2012a). Structure and substance in artificial-phonology learning: part I, structure. *Language and Linguistics Compass* **6**. 686–701.

Moreton, Elliott & Joe Pater (2012b). Structure and substance in artificial-phonology learning: part II, substance. *Language and Linguistics Compass* **6**. 702–718.

Moreton, Elliott, Joe Pater & Katya Pertsova (2015). Phonological concept learning. *Cognitive Science* .

Moreton, Elliott, Joe Pater & Katya Pertsova (2017). Phonological concept learning. *Cognitive Science* **41**. 4–69.

Moreton, Elliott & Katya Pertsova (2014). Pastry phonotactics: is phonological learning special? In Hsin-Lun Huang, Ethan Poole & Amanda Rysling (eds.) *Proceedings of the 43rd Annual Meeting of* *the Northeast Linguistic Society, City University of New York*, volume 2. Amherst, Massachusetts: Graduate Linguistics Students' Association, 1–14.

Moreton, Elliott & Katya Pertsova (2016). Implicit and explicit processes in phonotactic learning. In Jennifer Scott & Deb Waughtal (eds.) *Proceedings of the 40th Boston University Conference on Language Development*. Somerville, Mass.: Cascadilla, 277–290.

Moreton, Elliott, Brandon Prickett, Katya Pertsova, Josh Fennell, Joe Pater & Lisa Sanders (2021). Learning reduplication, but not syllable reversal. In Ryan Bennett, Richard Bibbs, Mykel Loren Brinkerhoff, Max J. Kaplan, Stephanie Rich, Nicholas Van Handel & Maya Wax Cavallaro (eds.) *Supplemental proceedings of the 2020 Annual Meeting on Phonology*. Washington, D.C.: Linguistic Society of America, ??–??

Morris, Peter E. (1981). The cognitive psychology of self-reports. In Charles Antaki (ed.) *The psychology of ordinary explanations of social behavior*, chapter 8. London: Academic Press, 183–204.

Munoz, Sergio R. & Shrikant I. Bangdiwala (1997). Interpretation of kappa and B statistics measures of agreement. *Journal of Applied Statistics* **24**. 105–112.

Muylle, Merel, Eleonore HM Smalle & Robert J Hartsuiker (2021). Rapid phonotactic constraint learning in ageing: evidence from speech errors. *Language, Cognition and Neuroscience* **36**. 746–757.

Myers, Scott & Jaye Padgett (2014). Domain generalisation in artificial language learning. *Phonology* **31**. 399–433.

Neisser, Ulrich & Paul Weene (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology* **64**. 640–645.

Newell, Allen & Paul S. Rosenbloom (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (ed.) *Cognitive skills and their acquisition*. Hillsdale, New Jersey: Erlbaum, 1–55.

Newell, B. R., J. C. Dunn & M. Kalish (2011). Systems of category learning: fact or fantasy? In B. Ross (ed.) *The psychology of learning and motivation*, volume 54, chapter 6. Academic Press, 167–215.

Newport, Elissa & Richard N. Aslin (2004). Learning at a distance i: statistical learning of non-adjacent dependencies. *Cognitive Psychology* **48**. 127–162.

Nisbett, R. E. & T. D. Wilson (1977). Telling more than we know: verbal reports on mental processes. *Psychological Review* **84**. 231–259.

Nosofsky, Robert M., Mark A. Gluck, Thomas J. Palmeri, Stephen C. McKinley & Paul Gauthier (1994a). Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition* **22**. 352–369.

Nosofsky, Robert M. & Thomas J. Palmeri (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin and Review* **3**. 222–226.

Nosofsky, Robert M., Thomas J. Palmeri & Stephen C. McKinley (1994b). Rule-plus-exception model of classification learning. *Psychological Review* **101**. 53–79.

Ohala, John J. (1993). The phonetics of sound change. In Charles Jones (ed.) *Historical linguistics: problems and perspectives.* Harlow: Longman, 237–278.

Onysko, Alexander, Marcus Callies & Eva Ogiermann (2013). Gender variation of anglicisms in German: the influence of cognitive factors and regional varieties. *Poznań Studies in Contemporary Linguistics* **49**. 103–136.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin and Review* **11**. 988–1010.

Paradis, Michel (2004). *A neurolinguistic theory of bilingualism.* Amsterdam: John Benjamins.

Pater, Joe & Elliott Moreton (2012). Structurally biased phonology: complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad* **3**. 1–44.

Pertsova, Katya (2012). Logical complexity in morphological learning. To appear in *Proceedings of the Berkeley Linguistics Society.*

Pertsova, Katya & Misha Becker (2020). In support of phonological bias in implicit learning. Manuscript submitted to *Language Learning and Development.*

Peterson, Margaret (1962). Some effects of the percentage of relevant cues and presentation methods on concept identification. *Journal of Experimental Psychology* **64**. 623–627.

Pycha, Anne, Pawel Nowak, Eurie Shin & Ryan Shosted (2003). Phonological rule-learing and its implications for a theory of vowel harmony. In M. Tsujimura & G. Garding (eds.) *Proceedings of the 22nd West Coast Conference on Formal Linguistics (WCCFL 22).* 101–114.

Rabi, Rahel & John Paul Minda (2016). Category learning in older adulthood: a study of the Shepard, Hovland, and Jenkins (1961) tasks. *Psychology and Aging* **31**. 185–197.

Rebei, Adnan, Nathaniel D Anderson & Gary S Dell (2019). Learning the phonotactics of button pushing: consolidation, retention, and syllable structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **45**. 2072.

Reber, A. S. (1993). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General* **118**. 219–235.

Rebuschat, Patrick (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning* **63**. 595–626.

Rescorla, R. A. & A. R. Wagner (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (eds.) *Classical conditioning*, volume II: Current research and theory. New York: Appleton–Century–Crofts, 64–69.

Saffran, Jenny R., Elissa L. Newport & Richard N. Aslin (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language* **35**. 606–621.

Sagey, Elizabeth (1990). *The representation of features in non-linear phonology: the Articulator Node Hierarchy.* New York: Garland.

Shepard, R. N., C. L. Hovland & H. M. Jenkins (1961). Learning and memorization of classifications. *Psychological Monographs* **75**.

Skoruppa, Katrin & Sharon Peperkamp (2011). Adaptation to novel accents: feature-based learning of context-sensitive phonological regularities. *Cognitive Science* **35**. 348–366.

Smalle, Eleonore HM, Merel Muylle, Arnaud Szmalec & Wouter Duyck (2017). The different time course of phonotactic constraint learning in children and adults: evidence from speech errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **43**. 1821.

Smith, J. David, Mark E. Berg, Robert G. Cook, Matthew S. Murphy, Matthew J. Crossley, Joe Boomer, Brian Spiering, Michael J. Beran, Barbara A. Church, F. Gregory Ashby & Randolph C. Grace (2012). Implicit and explicit categorization: a tale of four species. *Neuroscience and Biobehavioral Reviews* **36**. 2355–2369.

Smith, J. David, John Paul Minda & David A. Washburn (2004). Category learning in rhesus monkeys: a study of the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: General* **133**. 398–404.

Smith, J. David, Alexandrial Zakrzewski, Eric R. Herberger, Joseph Boomer, Jessica L. Roeder, F. Gregory Ashby & Barbara A. Church (2015). The time course of explicit and implicit categorization. *Attention, Perception, & Psychophysics* **77**. 2476–2490.

Smith, Nielson V. (1973). *The acquisition of phonology: a case study.* Cambridge, England: Cambridge University Press.

Sprouse, Jon (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgements in linguistic theory. *Behavior Research Methods* **43**. 155–167.

Taylor, Conrad F. & George Houghton (2005). Learning artificial phonotactic constraints: time course, durability, and relationship to natural constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **31**. 1398–1416.

Thaker, Pratiksha, Joshua B Tenenbaum & Samuel J Gershman (2017). Online learning of symbolic concepts. *Journal of Mathematical Psychology* **77**. 10–20.

Thomson, Ron I. & Tracey M. Derwing (2015). The effectiveness of L2 pronunciation instruction: a narrative review. *Applied Linguistics* **36**. 326–344.

Trapman, Mirjam & René Kager (2009). The acquisition of subset and superset phonotactic knowledge in a second language. *Language Acquisition* **16**. 178–221.

Tunney, Richard J & David R Shanks (2003). Subjective measures of awareness and implicit cognition. *Memory & cognition* **31**. 1060–1071.

Turk-Browne, Nicholas B., Phillip J. Isola, Brian J. Scholl & Teresa A. Treat (2008). The automaticity of visual statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **34**. 399–407.

Utulu, Don Chukwuemeka (2020). A description of some structures of (un)derived tones in èwúlú nouns. *European Journal of Literature, Language and Linguistics Studies* **4**. 36–51.

Vigo, Ronaldo (2009). Categorical invariance and structural complexity in human concept learning. *Journal of Mathematical Psychology* **50**. 203–221.

Vigo, Ronaldo (2013). The GIST of concepts. *Cognition* **129**. 138–162.

Wattenmaker, William D., Heather L. McQuaid & Stephanie J. Schwertz (1995). Analogical versus rule-based classification. *Memory and Cognition* **23**. 495–509.

Wegscheid, D., R. Schertler & J. Hietaniemi (2015). `Time::HiRes`. Software package, distributed by `perl.org`.

White, Peter A. (1988). Knowing more about what we can tell: 'introspective access' and causal report accuracy 10 years later. *British Journal of Psychology* **79**. 13–45.

Williams, Rick L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics* **56**. 645–646.

Zager, Laura A. & George C. Verghese (2007). Caps and robbers: what can you expect? *The College Mathematics Journal* **38**. 185–191.

Zellers, Margaret, Brechtje Post & John Williams (2011). Implicit learning of lexical stress patterns. In *International Congress of Phonetic Sciences*. 2260–2263.

Zettersten, Martin & Gary Lupyan (2020). Finding categories through words: more nameable features improve category learning. *Cognition* **196**. 104–135.

Zhang, Jie & Yuwen Lai (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology* **27**. 153–201.

Zubin, David A. & Klaus-Michael Köpcke (1984). Sechs Prinzipien für die Genuszuweisung im Deutschen: ein Beitrag zur natürlichen Klassifikation. *Linguistische Berichte* **93**. 26–50.