Implicit and explicit processes in phonological concept learning

(ANONYMIZED)

Version of August 24, 2020.

*(*ANONYMIZED*).*

Keywords: phonotactic learning, concept learning, implicit learning, explicit learning, inductive bias, complexity

Abstract

Analogous inductive problems arise in linguistic and non-linguistic pattern learning, raising the question of whether learners solve them the same way. In particular, learners of non-linguistic concepts use both implicit (intuitive) and explicit (rational) processes, which differ as to their facilitating factors, behavioral signatures, inductive biases, and proposed model architectures (connectionist vs. rule-based). This study asks whether the same applies to phonotactic learning. Nine experiments ($N = 1337$ participants) collected generalization responses, learning curves, response times, and detailed debriefings. Training conditions were varied to elicit different degrees of implicit or explicit learning. Subjective self-report predicted objective measures of implicit vs. explicit learning. Implicit and explicit learners were found in every condition of every experiment, and many participants reported using multiple approaches. Naïve participants discovered phonetic features and invented names for them. Learning mode affected inductive bias in a surprising way: Although the connectionist-vs.-rule-based proposal predicts, and non-linguistic concept-learning studies have found, that explicit learning improves performance on biconditionals relative to family-resemblance patterns, learners of phonotactics did the exact opposite in two experiments, apparently due to the larger number of irrelevant features in phonological stimuli. We conclude that phonological learning, like non-linguistic learning, is served by implicit and explicit processes with different inductive biases, and that individuals differ widely in their approach to the learning task.

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 increasingly clear that many experiments, results, and models in phonological learning have close parallels in work on non-linguistic learning (Finley and Badecker 2010; Y. R. Lai 2012; Moore-Cantwell, Pater, Staubs, Zobel, and Sanders 2017; Moreton 2012; Moreton and Pater 2012a, 2012b; Moreton, Pater, and Pertsova 2015; Moreton and Pertsova 2016; Pater and 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. Laboratory studies of phonological learning have typically assumed a single learning process, used by all participants and identical to the one used in natural language acquisition. We ask instead whether phonological learning is like non-linguistic learning 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 (i.e., different sensitivity to different pattern types), and different algorithmic architectures.

To address this question, 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 to compare that with objective measures of performance. The other is evaluation, not just of end-state performance, but of how performance changes over time, in order to compare it with the predictions of different learning models.

2 Implicit and explicit learning

Studies of inductive learning of non-linguistic patterns (aka "concepts" or "categories") have led many psychologists to hypothesize two concurrent learning processes for non-linguistic patterns, 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; J. D. Smith et al., 2012, 2015). The two systems correspond approximately to the familiar notions of *reasoning* and *intuition*. Each is characterized by a set of putatively co-occurring properties.

The implicit system is thought to be effortless, unconscious, and undemanding of attention or working memory. Architecturally it is proposed to be "cue-based", i.e., learning can be modelled as incremental weight update on an array of property detectors. Hence, learning is gradual rather than abrupt, closed to conscious introspection, and faster for patterns which are supported by multiple overlapping cues than for

those that are supported by a small number of disjoint cues. The explicit system is hypothesized to be effortful, conscious, 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 and Trabasso 1964), open to introspection, and faster for patterns which depend on fewer features. Several variants of this two-systems hypothesis exist; for critical reviews see Evans (2008); Keren and Schul (2009); B. R. Newell, Dunn, and Kalish (2011); Osman (2004).[1]

The use of each of these two systems is predicted to produce a recognizable syndrome of behavioral effects. Since the explicit system is conscious and effortful, participants ought 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 verbalize the target rule. In an experiment where a partly-correct rule is no help, explicit learners should fall into two groups at the end of training: those who achieve a high level of classification accuracy and are able to accurately verbalize the target rule, and those who are near chance and state an inaccurate rule or no rule. If trial-by-trial classification 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 switch from rule-seeking to rule-using.

In an implicit learner, on the other hand, the product of learning is a non-verbalizable set of continuous weights on an array of property detectors; hence, implicit learners should be unable to accurately verbalize the target rule. Since the weights are updated incrementally and automatically, 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 hypothesized to bias it in favor of rules that involve simple logical relations between a small number of features, such as two-feature exclusive-or patterns ("exactly one of green or square"), 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 ("differs by at most one feature from a small green square"). Empirical evidence for the occurrence of these symptoms in non-linguistic learning is summarized in Table 1.

Different experimental conditions have been found to 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 verbalize favor implicit learning (Table 2).

[1] The phonological knowledge that we are studying does not fit comfortably into the declarative-procedural schema, since it is knowledge of generalizations about the shape of words rather than about how they combine, and because it can be implicit, but is not obviously procedural. We thus focus on the implicit-explicit distinction, as defined by (e.g.) (Lee, 1995; Mathews et al., 1989; Reber, 1993; J. D. Smith et al., 2015), rather than procedural vs. declarative.

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, and Austin (1956); Ciborowski and Cole (1972)
Can state correct rule	yes	no	Ciborowski and Cole (1973)
Correctness of stated rule predicts performance	yes	no	Lindahl (1964)
Shape of learning curve	abrupt	gradual	J. D. Smith, Minda, and Washburn (2004)
Progression of RTs	abrupt	gradual	Haider and Rose (2007)
Distribution of test-phase performance	bimodal	unimodal	Kurtz, Levering, Stanton, Romero, and Morris (2013)
Structural bias	IFF/XOR easier than family-resemblance	IFF/XOR advantage reduced or reversed	Kurtz et al. (2013); Love (2002); Rabi and 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 and Markman (2003)
Intent	intentional	incidental	Love (2002)
Features	verbalizable	not verbalizable	Kurtz et al. (2013); Nosofsky and Palmeri (1996)

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 separability of their physical instantiations, might affect learning (Kurtz et al., 2013; Minda, Desroches, & Church, 2008; Nosofsky & Palmeri, 1996; Zettersten & Lupyan, 2020), the learning processes themselves are proposed to be general-purpose problem solvers. It follows that both processes ought to be applicable to the domain of language, and indeed, both implicit and explicit processes have been found to be involved in

language learning (Ellis 1994; for a recent review see K. Lichtman 2013).[2] 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-Vrooman, 1990; DeKeyser, 2003; Paradis, 2004). However, 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 (K. M. Lichtman, 2012; Reber, 1993).

There has been little, if any, study contrasting implicit vs. explicit learning of natural first- or second-language phonology.[3] Studies of phonological learning in artificial languages are predominantly 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. Even overt criticisms of "artificial-language" methodology as contaminated by explicit learning (e.g., Zhang and Lai 2010) have not presented evidence that it actually is so contaminated, and have not led to attempts to remedy the problem. Experimenters may design their experiments to minimize explicit learning (e.g., Do, Zsiga, and Haverhill 2016; Glewwe 2019), or exclude data from participants who correctly verbalize the pattern, but, with some recent exceptions (Kimper 2016; Moreton and Pertsova 2016), they do not normally analyze implicit and explicit learners separately, or distinguish implicit learners from failed explicit learners.

Our current lack of knowledge about the algorithmic diversity of phonological learning is a serious 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 therefore asks whether the inductive learning of phonology in the lab is served by implicit and explicit processes that are like the ones proposed for non-linguistic inductive 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. Experiments 1–5 focus on identifying correlates of implicit vs. explicit learning modes using single-feature assertions ("Type I" patterns, in the terminology of Shepard, Hovland, and Jenkins 1961).

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

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

Experiments 6–9 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.[4] The experiments are summarized in Tables 3 and 4.

This study goes beyond previous work on phonological 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 (Table 1) in a single study.

	Experiment									
	1		2		3		4		5	
	E-P	I-P	E-P	I-P	E-P	I-P	E-P	I-P	E-P	I-P
Valid participants	74	63	26	22	43	38	43	39	99	77
Training										
Feedback	Y	N	Y	N	Y	N	Y	N	Y	Y
Rule instructions	Y	N	Y	N	Y	N	Y	N	N	N
Stimuli per trial	2	1	2	1	2	1	1	1	2	2
	$(+,-)$	$(+)$	$(+,-)$	$(+)$	$(+,-)$	$(+)$	$(+/-)$	$(+)$	$(+,-)$	$(+,+)$
Learning scenario	gender		gender		vocabulary		gender		vocabulary	
Features										
2/3 syllables	Y		Y		Y		Y		Y	
1st/2nd stress	Y				Y		Y		Y	
C's same/different	Y									
V front/back	Y									
Fricatives/stops	Y		Y		Y		Y		Y	
Labial/coronal	Y		Y							
Hypotheses										
1. Explicit-promoting facilitates										
Rule-Seeking	supported		not found		not found		not found		not found	
and Rule-Stating	supported		supported		supported		not found		supported	
2. Rule-Seeking facilitates										
Stating	supported		supported		supported		supported		supported	
and Rule Correctness	supported		supported		supported		supported		supported (E-P onl	
3. Rule Correctness facilitates generalization	supported		supported		supported		supported		supported	
4. Rule Correctness is associated with abrupt learning curve	supported		supported		not found		not found		supported	
5. Rule Correctness is associated with RT acceleration at last error	supported		not found		not found		not found		supported	

Table 3: Summary of Experiments 1–5, focusing on single-feature ("Type I") patterns. "+" and "−" represent positive (pattern-conforming) and negative (nonconforming) stimuli.

[4]The experiment numbers do not reflect the historical sequence. Experiments 2 and 3 were run simultaneously with Experiments 6 and 8, respectively, but are presented separately with the other Type I results in order to make the article clearer and shorter.

	Experiment							
	6		7		8		9	
	E-P	I-P	E-P	I-P	E-P	I-P	E-P	I-P
Valid participants	53	35	75	76	75	58	55	64
Training								
Feedback	Y	N	Y	N	Y	N	Y	Y
Rule instructions	Y	N	Y	N	Y	N	N	N
Stimuli per trial	2	1	2	1	2	1	2	2
	$(+,-)$	$(+)$	$(+,-)$	$(+)$	$(+,-)$	$(+)$	$(+,-)$	$(+,+)$
Learning scenario	gender		gender		vocabulary		vocabulary	
Features								
2/3 syllables	Y		Y		Y		Y	
1st/2nd stress			Y		Y		Y	
C's same/different								
V front/back								
Fricatives/stops	Y		Y		Y		Y	
Labial/coronal	Y							
Hypotheses								
1. Explicit-promoting facilitates								
Rule-Seeking	not found		not found		supported		not found	
and Rule-Stating	supported		supported		not found		not found	
2. Rule-Seeking facilitates								
Stating	supported		supported		supported		supported	
and Rule Correctness	—		supported		supported		not found	
3. Rule Correctness facilitates generalization	supported (E-P)		not found		supported		supported	
4. Rule Correctness is associated with abrupt learning curve	not found		not found		not found		not found	
5. Rule Correctness is associated with RT acceleration at last error	not found		not found		not found		not found	
6. Rule-Seeking facilitates II over IV	contradicted		contradicted		not found		not found	

Table 4: Summary of conditions and results from Experiments 6–9, focusing on biconditional ("Type II") vs. family-resemblance ("Type IV") patterns. "+" and "−" represent positive (pattern-conforming) and negative (nonconforming) stimuli.

3 Experiment 1

Experiment 1 was a straightforward test of the hypotheses described in Section 2 above in a linguistic context: The conditions in Table 2 were varied to see if they had the effects in Table 1.[5] The Implicit-Promoting condition was based on a widely-used phonotactic-learning paradigm in which participants are familiarized 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, 2010, 2016; Gerken, Quam, and Goffman 2019; Greenwood 2016; Kuo 2009; R. Lai 2015; Moreton 2008, 2012; Moreton, Pater, and Pertsova 2017; Skoruppa and Peperkamp 2011). The Explicit-Promoting condition differed in that training trials consisted of choosing the conforming member of a conforming-non-conforming pair.

[5]Parts of this experiment were previously presented as (ANONYMIZED).

Participants in both conditions of Experiment 1 were told that they would be learning to distinguish words of the target gender from words of another gender.[6] Many natural languages assign gender 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 (Franco, Zenner, & Speelman, 2018; Onysko, Callies, & Ogiermann, 2013; Zubin & Köpke, 1984). In this experiment, each participant's "language" assigned nouns feminine or masculine gender based on a visual or phonological feature chosen randomly from a larger array. Participants were exposed to a training set under conditions hypothesized to favor either implicit or explicit learning (see Section 2), and were then tested on generalization of the pattern by classifying novel stimuli as feminine or masculine. A post-experiment questionnaire was used to assess self-reported rule-seeking and correctness of any stated rule.

Experiment 1 is the first in a series of five similar experiments in which the pattern depends on a single feature (Table 3, above). The experimental and analytic procedures for Experiment 1 are described here in detail. Only the differences of subsequent experiments from Experiment 1 will be spelled out.

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 shown in Table 5.

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
$\acute{\sigma} = \sigma_1$	VCəC	VCəCəC
$\acute{\sigma} = \sigma_2$	əCVC	əCVCəC

Table 5: Schema used to construct the auditory nonword stimuli for all experiments.

Six phonological variables were chosen based on the authors' expectations that each would be individually

[6] In the above-cited experiments corresponding to the Implicit-Promoting condition (Carpenter 2006, etc.), the familiarization task was explained to participants as listening to "words" in a "language", and the test task as distinguishing new words 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. The gender task was used instead to improve ecological validity.

[7] Vowel-initial words were used for forward compatibility with other planned experiments (not reported here). Similar patterns are attested in a number of natural languages; e.g., in Urama (Papua New Guinea; Trans-New Guinea) all verbs begin with a vowel (J. Brown, Muir, Craig, & Anea, 2016; J. L. Brown, 2009). In Ẹdo and its relative Urhobo (both Nigeria; Niger-Congo, Edoid), all nouns, or nearly all, begin with a vowel (Ọmọruyi, 1986; Kelly, 1969). In Arrernde (Arrernte, Aranda; Australia, Pama-Nyungan), all words are vowel-initial at the surface level (Breen & Pensalfini, 1999).

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 verbalize: two vs. three syllables, first- vs. second-syllable stress, all consonants different vs. all consonants identical. The other three were chosen with the opposite expectation: 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 were 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 normalized 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 in the test phase), leaving 191 valid participants. In addition to the six phonological-feature conditions described above, there were also three visual-feature conditions, discussed in a separate publication (ANONYMIZED). 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.

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 36 groups defined by crossing Training Group (Explicit-Promoting vs. Implicit-Promoting) with Critical Feature (any of the six phonological and three visual features) and Target Gender (feminine vs. masculine).

A "language" was generated for the participant as follows. If the participant was in a phonological Critical Feature group, that feature and the two others in its Critical Feature subgroup (the intended-verbalizable and intended-nonverbalizable subgroups) were chosen as the axes of the three-dimensional stimulus space. Eight bins were created, corresponding to the eight combinations of the two values on each of the three axes. Four of the bins thus corresponded to pattern-conforming feature values, and four to non-conforming feature values. The set of all 512 nine-digit binary numbers (representing all possible combinations of the nine phonological and visual features) was randomly shuffled, and numbers were drawn and assigned to bins based on the Critical Feature number until the eight bins each had sixteen numbers. Each number was then converted to a word-plus-picture pair by randomly choosing a word that matched the phonological features and a picture that matched the visual ones. One value of the binary Critical Feature was randomly chosen to be the pattern-conforming value, and the half of the word-plus-picture pairs that had that feature were assigned the Target Gender, while the other half were assigned the other gender. The resulting 128 word-plus-picture-plus-gender triplets were randomly divided into 32 conforming and 32 non-conforming items for the training phase, and another 32 and 32 for the test phase. In the written instructions, 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.

The *Implicit-Promoting* training condition was designed to encourage implicit learning (see Section 2 above). Participants in this group were instructed that they would be learning words in the artificial language, and that all of the words they were to learn would be feminine (or masculine, depending on the Target Gender group for that participant). On each training trial, the participant saw one of the pictures, captioned with the English word for that picture, and a single gray mouse button below it (Figure 1, left panel). Mousing over the button caused the correct word for that picture in the artificial language to be played (as often as desired). Pressing the button triggered the next trial after a 250-ms delay. Only the 32 pattern-conforming training stimuli were presented. All 32 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.

The *Explicit-Promoting* training condition was designed to encourage explicit learning (see Section 2 above). Participants in this group were instructed that they would learn to tell whether a word was feminine or masculine, by trial and error; that there were systematic differences between the feminine and masculine

words; and that by recognizing the systematic differences, the participant could get the right answer every time. On each training trial, participants saw two pictures, each with a button below it (Figure 1, right panel). Mousing over the button played the name of the picture (as often as desired). 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. For an incorrect response, the feedback was a sad two-note trumpet measure, after which the software did not advance to the next trial, but waited for the participant to click on the correct button. The 32 positive and 32 negative stimuli were randomly paired and ordered to produce four-trial blocks in which each of the four pattern-conforming bins and each of the four non-conforming bins occurred once. After all 32 conforming-nonconforming pairs had been presented, they were re-paired, reordered, and re-presented, until they had been presented four times, or until the participant had responded 100% correctly on four consecutive blocks ("reached criterion").

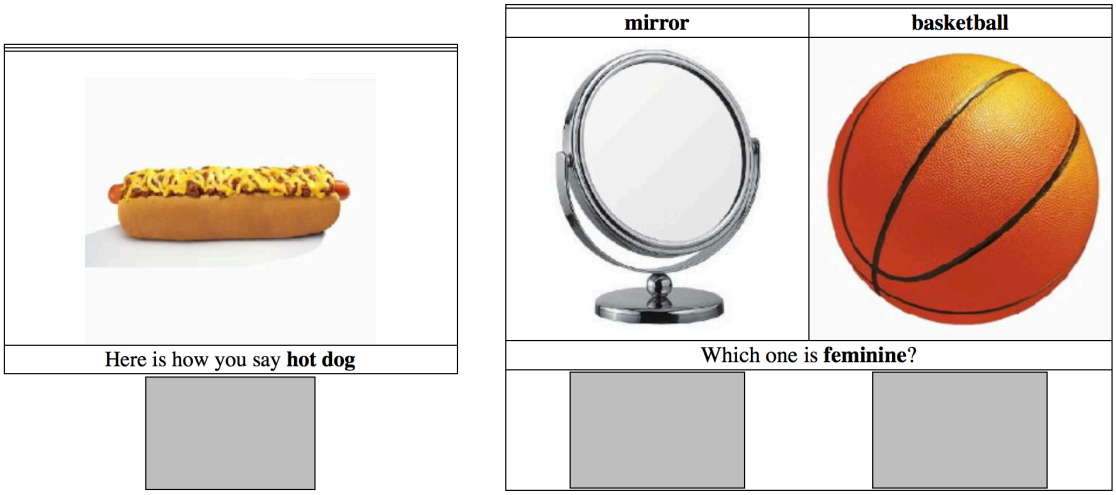


Figure 1: Participant view of a trial in Experiment 1. Left panel: Training phase, Implicit-Promoting condition. Right panel: Training phase, Explicit-Promoting condition, and test phase, both conditions.

In both training conditions, participants 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 a correct answer was received by the server. This was done using the `time` function in the `Time::HiRes` module in Perl (Wegscheid, Schertler, & Hietaniemi, 2015). Since response times were measured at the server, they 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, all of which add variability to the durations (Høiland-Jørgensen, Ahlgren, Hurtig, & Brunstrom, 2016).

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

Participants in both training conditions were notified after the 64th training trial that they had completed "at least" half of the training. 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 instead of the conforming and non-conforming training items, 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 (Figure 1, right panel).

The experiment was followed by a debriefing questionnaire. In addition to demographic 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 6.

3.1.3 Questionnaire coding

Each participant's questionnaire responses were coded according to the following criteria (for the full coding rubric, see Appendix):

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. Every response was coded by both scorers. To minimize 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.[8] Responses to the open-ended training-phase and test-phase strategy questions were displayed to the experimenter simultaneously, together with a statement of the correct rule, but without information as to what Training Group the participant was in. An example of the display format seen by the experimenters during coding is shown in Figure 2. 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.

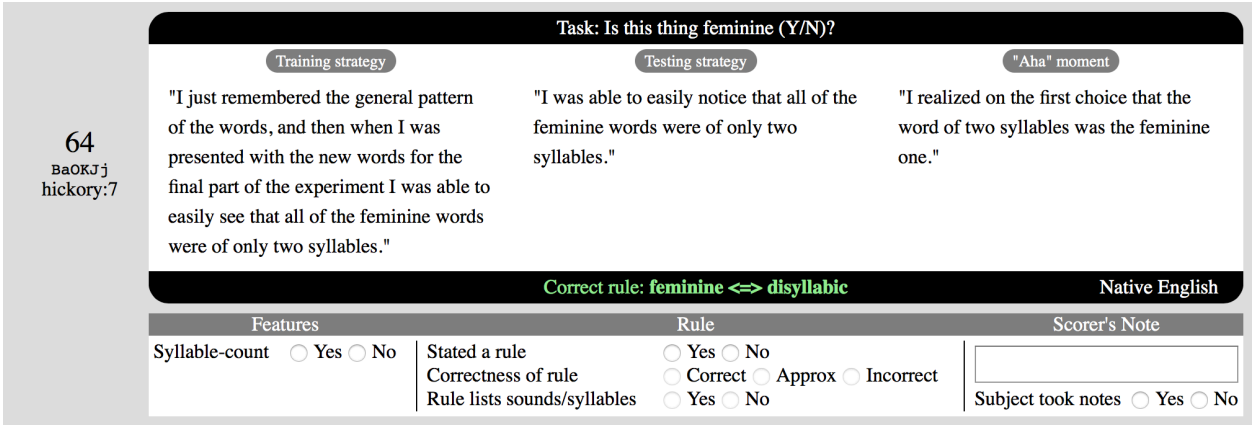


Figure 2: Display format for questionnaire coding. ("BaOKJj hickory:7" is a lab-internal code.)

[8]Each experiment's questionnaire data was originally coded by two coders, not necessarily the same ones, as soon as the experiment had been run, and these codes were used for interim analyses. For this publication, the entire data set was combined and re-scored in the interests of uniformity across experiments.

Questionnaires from 1337 participants were scored. This number includes participants whose data was later excluded from analysis (e.g., for reporting a non-English first language). Inter-rater disagreements are tabulated in Table 7. 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), as shown in Table 7. 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).

Variable	Disagreements	Cohen's κ
Feature stating	79	0.838
Rule stating	111	0.818
Rule correctness	54	0.804
Listing	25	0.880
N	1337	

Table 7: Number of scorer disagreements for each of the scored variables. This table includes all participants that were scored, including those whose data was later excluded from analysis.

3.2 Hypotheses and planned analyses

If 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.

Memorizer: Checked box "Tried to memorize 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).

These categories were not mutually exclusive. Many participants reported switching between approaches during the experiment, e.g.,

> I tried to find a rule where the words sounded like ah or round items ended with pup or something similar. I couldn't really find a pattern so I just started to memorize the words as best as I could. (Participant zqdzVh, Experiment 5 condition)

> At first I tried to memorize the words because I thought that might be useful, but I soon realized that looking for a rule would be more effective. The rule I found very quickly was that the correct word was always three syllables, starting with a vowel and then two syllables starting with consonants. Following that, I just selected the words that fit that pattern. (Participant eBvXUY, Experiment 5 condition)

> at first I tried to memorize the words but when I saw there were new words I used my intuition instead (Participant qspwQc, Experiment 8 condition)

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 (Table 2), then signatures of explicit learning (Table 1) should occur with greater frequency in the Explicit-Promoting condition than in the Implicit-Promoting condition. In particular, a greater proportion of Explicit-Promoting participants should be Rule-Seekers and Rule-Staters (*Hypothesis 1*). If the explicit system is indeed under voluntary control, then the products of that system (namely rules) ought to be reported more often by participants who report voluntary use of that system. I.e., Rule-Seekers should be more likely than others to be Rule-Staters (*Hypothesis 2*).

Self-report of cognitive processes is often viewed skeptically, and not without reason, as there are numerous instances in which self-report proves to be unrelated, or only coincidentally related, to objective measures of performance (Nisbett & Wilson, 1977), including learned problem-solving performance (Berry & Broadbent, 1984). However, that does not justify dismissing self-report out of hand. Self-report is frequently corroborated by behavior, especially in intentional problem-solving tasks (Ericsson & Simon, 1980; Kellogg, 1982; Morris, 1981; White, 1988), and one goal of this experiment series is to test the validity of self-report in phonological learning. Accurate introspection into the explicit system implies that if a participant states an explicit rule, that rule should 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. Some may respond on the basis of an approximately-correct explicit rule, stated or unstated, with performance better than chance but worse than perfect. Others may use an incorrect explicit rule, stated or unstated, based on an irrelevant feature (e.g., smooth vs. rough texture), and are expected to perform at 50%. Others may respond on the basis of an intuitive familiarity with the pattern, acquired via gradual cue-based learning (see Section 2), and hence

may show above-chance preference for pattern-conforming items. Still others may have met criterion by memorizing the training stimuli, and therefore be at chance when confronted with novel test stimuli. Finally, there may be some participants who found and used the correct explicit rule, but omitted to say so on the questionnaire; their performance should be near 100%. Since there is no certain way to separate these subgroups, the most we can say is that those who did not state a correct or approximately-correct rule should show a wide distribution of somewhat above-chance performance. In any case, the more correct the stated rule is, the higher the performance should be on the novel test stimuli (*Hypothesis 3*).

By comparing Solvers with each other, we can compare participants who achieved the same level of 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. Hence, among Solvers, Correct Staters are predicted to be more likely than other Solvers to show abrupt improvement in 2AFC performance (*Hypothesis 4*) and a decrease in response times (*Hypothesis 5*) 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 commonly assumed to occur in a phonological-learning experiment and what our participants reported, we quote their own words before proceeding to a quantiative analysis.

Naïve participants, i.e., those who self-reported not having studied linguistics, were able to discover phonetic properties and invent ways to verbalize them. This was true even for properties which take time and effort for many Linguistics 101 students to grasp. The continuancy distinction (fricatives vs. stops) was intended by the experimenters to be non-verbalizable, but many participants recognized 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, stops/fricatives condition)

> Thought I found a weak pattern with hard endings (-t, -b, -d, etc) but there were some soft endings (-f, -s) that threw me off (Participant HaLObc, edible/inedible condition)

> I tried to identify what sounds were consistent. Words ending in softer sounds tended to be masculine. I looked for soft ending sounds. Some words sounded like they came from the first set. (Participant BHfSgt, fricatives/stops condition)

> 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 pzyaXQ, fricatives/stops condition)

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 condition)

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

> Found a rule, ud uz us ut are all feminine, ub uf up uv are all masculine (Participant DChrth, labial/coronal condition)

> 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 condition)

> I listened to how the last consonant was pronounced and looked for a rule. The words ending with d, t, s, or z were masculine. If the word ended in a d, t, s, or z I choose that as the masculine word. (Participant MmjUXR, labial/coronal condition)

> 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 condition)

> I looked for the sound of the vowels, like the u or o sound. (Participant OBDBYj, front/back vowel condition)

Instead of three easily-verbalizable and three non-verbalizable features, as intendent, the experiment turned out to have used one feature that was frequently verbalized as a feature (two vs. three syllables), two features that were frequently verbalized as letter lists (fricatives vs. stops and labials vs. coronals), one

feature that was frequently verbalized ambiguously as a feature or a letter (initial vs. second-syllable stress; see Section 3.1.3), and two that were rarely verbalized (same vs. different consonants and front vs. back vowel). Summary statistics are shown in Table 8.

	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 8: 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 Appendix.

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 9). 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:

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 condition)

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 condition)

Intuition and sequential hypothesis testing:

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

Table 9: Self-reported learning strategies (check-box responses), Experiment 1.

> 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 condition)

Tried intuition but discovered rule:

> 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 condition)

Memorization:

> 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 condition)

Tried rule-seeking but switched to memorization:

> 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 condition)

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 condition)

The reports differ from one participant to the next, even within a single condition, giving at least an initial impression that when an experiment samples participants, it samples 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.

3.3.2 Hypothesis 1: Effect of training condition on rule-seeking and -stating

Results from all participants are plotted in Figure 3. It is apparent that participants in the Explicit-Promoting condition, who were instructed to seek a rule and given right-wrong feedback on every trial, were indeed significantly more likely than those in the Implicit-Promoting condition to be Rule-Seekers and Rule-Staters (Table 10, $p = 0.0001643$ and 0.01053 respectively by Fisher's exact test, two-sided).

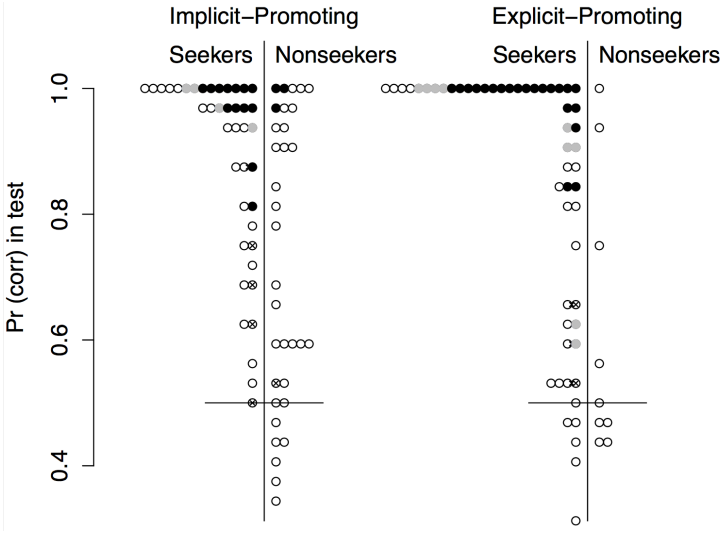


Figure 3: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 1. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance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

	Rule-Seeker		Rule-Stater	
Training Condition	T	F	T	F
Explicit-Promoting	54	9	37	26
Implicit-Promoting	41	33	27	47

Table 10: Rule-Seeking and Rule-Stating as a function of Training Condition, Experiment 1

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

Coefficient	Estimate	Std. Error	χ^2 value	$\Pr(>\|z\|)$	
(Intercept)	-2.9444	1.5294	9.8926	0.0017	**
Seeker	3.7065	1.5570	15.4642	0.000084	***
Implicit-Promoting	1.0641	1.6134	0.6094	0.435	
Seeker × Implicit-Promoting	-1.5870	1.6695	1.3206	0.250	

Table 12: Fitted Firth logistic-regression model for Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 1

3.3.3 Hypothesis 2: Effect of rule-seeking on rule-stating and rule correctness

It is also apparent from Figure 3 that in both training conditions, Seekers were much more likely than Non-Seekers to be Staters (Table 11). Because some cells were empty or nearly so (in this and subsequent experiments), a Firth-penalized logistic-regression model was fit using the `logistf` method in R's `logistf` package (Firth, 1993; Heinze & Ploner, 2018), with Stated (0 or 1) as the dependent variable and Seeker (0 or 1), Implicit-Promoting (0 or 1), and their interaction as independent variables. In the fitted model (Table 12), the significant effect of Seeker shows that Rule-Seekers were significantly more likely to be Rule-Staters, and the small size and non-significance of the coefficients for Implicit-Promoting and Sought × Implicit-Promoting confirm that this held in both Training Groups. Likewise, Seekers were much more likely than Non-Seekers to be Correct or Approximate Staters (Table 11; analogous logistic-regression model in Table 13).

Coefficient	Estimate	Std. Error	χ^2 value	$\Pr(>\|z\|)$	
(Intercept)	-2.9444	1.5294	9.8926	0.0017	**
Seeker	3.1634	1.5537	10.4099	0.0013	**
Implicit-Promoting	0.7794	1.6331	0.2958	0.5865	
Seeker × Implicit-Promoting	-1.3350	1.6860	0.8650	0.3523	

Table 13: Fitted Firth logistic-regression model for Correct and Approximate Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 1

3.3.4 Hypothesis 3: Effect of rule correctness on generalization

Figure 3 also shows that participants tend to fall into two groups: Correct or Approximately-Correct Staters, who perform nearly perfectly on the generalization 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. That is clearly consistent with Hypothesis 3.

The effect of rule discovery on generalization performance was quantified using *plan logistic regression* with a two-stage sampling model. This procedure, also known as a "complex survey design", "sampler's model", or "population average 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 (an individual voter). Plan logistic regression is an alternative way of taking into account within-participant dependency (Bieler & Williams, 1995; Williams, 2000) while avoiding convergence problems encountered when trying to fit mixed-effects logistic regression models to individual 2AFC responses in some of the experiments in this paper. (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.) The models were fit using the R package `survey` (Lumley, 2004, 2019; 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 14. The significant 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	

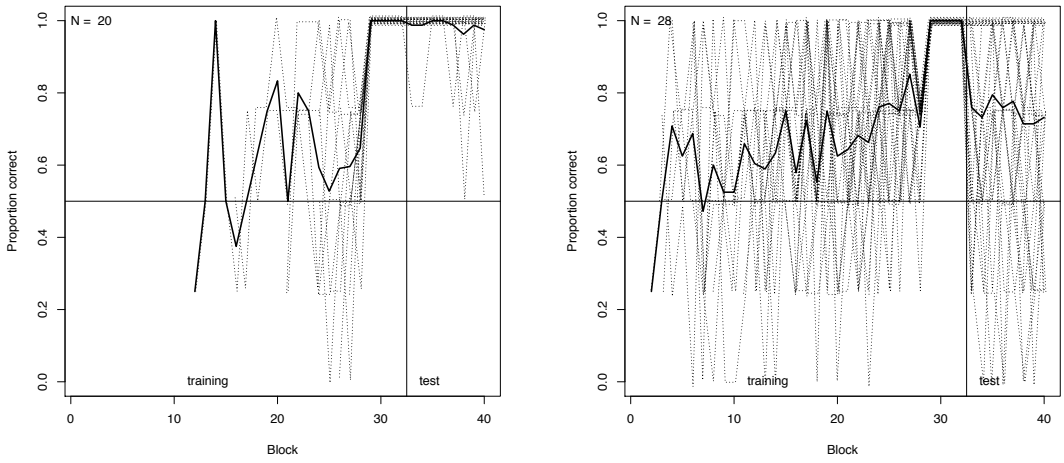
Table 14: Summary of plan logistic-regression model for pattern-conformity of generalization-test responses, Experiment 1 (4384 responses from 137 participants).

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 15: Summary of the the 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.)

3.3.5 Hypotheses 4 and 5: Effect of correct rule-stating on abruptness and response time

The classification task in 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 16 consecutive correct trials before the end of the experiment) 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 15 (fitted using `svyglm`, as above). This is as predicted by Hypothesis 4: 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 3, above).



Correct and Approximate Staters | Others

Figure 4: Learning curves for Solvers in the Explicit-Promoting condition of Experiment 1, aligned to last error. Dashed lines are individuals, solid line is the mean across participants.

Hypothesis 5 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-trial radius. Trial durations of less than 4 seconds or more than 30 seconds were excluded. Both 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 (A. 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. A general linear model was then fit via `svyglm` as above, 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.

The fitted model is shown in Table 16. 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 and Rose (2007), in which rule discovery enables the participant to respond correctly after listening to only one of the two stimuli.

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 16: 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).

3.4 Discussion

Participants in both conditions reported rule-seeking and stated rules, but did so more often in the Explicit-Promoting condition. In both conditions, (subjective) self-report of rule-seeking was associated with a higher (objective) rate of rule-stating. Generalization was better than chance in both training conditions and regardless of rule-seeking, but the more correct the stated rule was, the more pattern-conforming choices the participant made in the generalization test. Correct rule-stating was associated with perfect or near-perfect generalization performance. In the Explicit-Promoting condition, Solvers who later stated a correct rule showed an abrupt performance jump and acceleration in response time at the last pre-criterion error. These results support the hypothesis that participants can learn the phonotactic pattern using both implicit and explicit processes (Section 2), and that explicit learning is possible even when the relevant features may not appear *a priori* verbalizable. These results also confirm that participants' self-report of their mental proceses is at least partly accurate.

4 Experiment 2

Experiment 1 provided hard evidence as to how verbalizable the phonological features really were (Table 8). Experiment 2 attempts to replicate Experiment 1, reducing between-participant variance by focusing on three features which led to high rates of correct or approximately-correct rule-stating among Seekers in Experiment 1 (Table 8).

4.1 Methods

The critical features included two/three syllables, fricatives/stops, and labials/coronals. The participant pool, procedure and the post-experiment questionnaire were identical to the ones used in Experiment 1. Of the 55 participants who completed the experiment, 7 were excluded from analysis (2 reported a non-English L1, 4 reported taking written notes, 1 reported choosing test-phase responses that were maximally *un*like what they were trained on), leaving 48 valid participants.

4.2 Results

Results were analyzed as in Experiment 1. Descriptions of the analysis procedure will therefore be abbreviated in this and subsequent experiments, except where they differ from the corresponding procedures in Experiment 1.

4.2.1 Hypothesis 1: Effect of training condition on rule-seeking and -stating

Results are plotted in Figure 5. Participants in the Implicit-Promoting condition were numerically less likely than those in the Explicit-Promoting condition to be Rule-Seekers, but the difference was not significant by Fisher's exact test (two-sided, $p = 0.3408$). Participants in the Implicit-Promoting condition were significantly less likely than those in the Explicit-Promoting condition to be Rule-Staters ($p = 0.0216$).

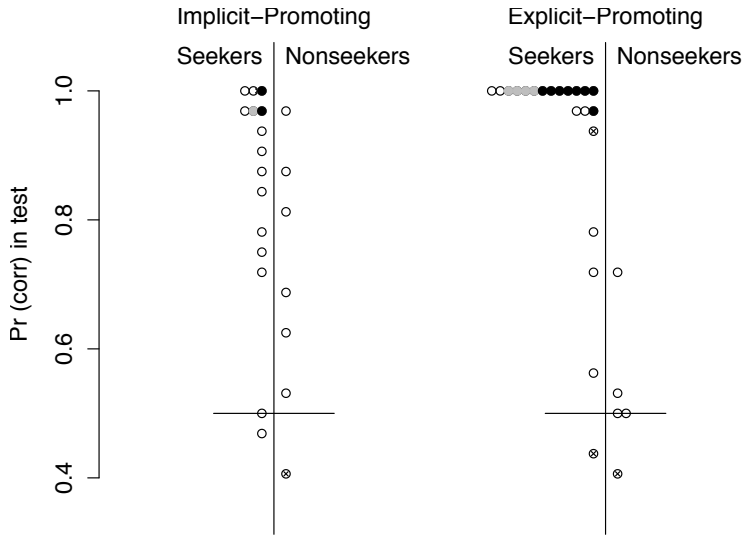


Figure 5: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 2. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

	Rule-Seeker		Rule-Stater	
Training Condition	T	F	T	F
Explicit-Promoting	21	5	16	10
Implicit-Promoting	15	7	5	17

Table 17: Rule-Seeking and Rule-Stating as a function of Training Condition, Experiment 2

4.2.2 Hypothesis 2: Effect of rule-seeking on rule-stating and rule correctness

Figure 5 Table 18 shows that two-thirds of the Seekers in the Explicit-Promoting condition (41/62) were Staters, as opposed to a quarter of all other participants (19/74). Seekers were again significantly more likely than Non-Seekers to be Staters, but no significant effect of, nor interaction with, Training Condition was found (Table 19). Seekers in the Explicit-Promoting condition were also significantly more likely than Non-Seekers to be Correct or Approximate Staters, and those in the Implicit-Promoting condition did not differ significantly from them (Table 20).

	Explicit-Promoting		Implicit-Promoting	
	Seekers	Non-Seekers	Seekers	Non-Seekers
Non-Staters	6	4	11	6
Staters	15	1	4	1
Correct Staters	8	0	2	0
Approximate Staters	4	0	1	0
Incorrect Staters	3	1	1	1

Table 18: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 2

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	-1.0986	1.0327	1.5697	0.2102	
Seeker	1.9676	1.1381	4.0660	0.0437	*
Implicit-Promoting	-0.3677	1.4157	0.0789	0.7786	
Seeker × Implicit-Promoting	-1.4395	1.6009	0.9116	0.3396	

Table 19: Fitted Firth logistic-regression model for Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 2

Coefficient	Estimate	Std. Error	χ^2 value	$\Pr(>\|z\|)$	
(Intercept)	-2.3978	1.6180	4.8757	0.0272	*
Seeker	2.6723	1.6769	5.1431	0.0233	*
Implicit-Promoting	-0.3101	2.2486	0.0222	0.8814	
Seeker × Implicit-Promoting	-1.2372	2.3749	0.3041	0.5813	

Table 20: Fitted Firth logistic-regression model for Correct and Approximate Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 2

4.2.3 Hypothesis 3: Effect of rule correctness on generalization

Pattern-conforming responses were coded as 1, non-conforming responses as 0. The fitted model is shown in Table 21. The large and significant coefficient for *Rule Correctness* indicates that more correct the stated rule was, the more likely the participant was to give pattern-conforming test-phase responses.

Coefficient	Estimate	Std. Error	t value	$\Pr(>\|z\|)$	
(Intercept)	0.9369	0.2913	3.217	0.00243	**
Rule Correctness	6.6844	2.0371	3.281	0.00203	**
Implicit-Promoting	0.2824	0.3787	0.746	0.45978	
Rule Correctness × Implicit-Promoting	−3.3386	2.2109	−1.510	0.13817	

Table 21: Summary of fitted logistic-regression model for pattern-conformity of generalization-test responses, Experiment 2. (1536 responses from 48 participants.)

4.2.4 Hypotheses 4 and 5: Effect of correct rule-stating on abruptness and response time

The 16 trials preceding the last error for Solvers in the Explicit-Promoting condition were analyzed as in Experiment 1. Solvers who were not Correct or Approximate Staters were nonetheless performing significantly above chance before their last error. The coefficient for Rule Correctness is large and significantly negative, indicating that correctness of the stated rule was associated with less-accurate performance preceding the last error — i.e., with a more abrupt transition to flawless performance.

The response-time acceleration across the last error that was found in Experiment 1 (Table 16) failed to replicate in this experiment: The coefficient for *Preceding* × *Rule Correctness* was near zero and non-significant ($\beta = -0.0155, p = 0.854$).

4.3 Discussion

Most of the results of Experiment 1 were replicated in Experiment 2. Rule-seeking was common in both the Implicit- and Explicit-Promoting conditions. Although the Explicit-Promoting training condition did not facilitate rule-seeking this time relative to the Implicit-Promoting condition, it did facilitate rule-stating.

Coefficient	Estimate	Std. Error	z value	t value	$\Pr > \|t\|$
(Intercept)	2.1619	0.2259	9.570	$5.15e-06$	***
Rule Correctness	−1.5796	0.5576	−2.833	0.0196	*

Table 22: Summary of the the 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. (126 responses from 11 participants, excluding 7 more participants who either never made an error, or who only made an error on their first trial.)

Rule-seeking facilitated rule-stating and rule correctness, and rule correctness improved generalization performance and increased the abruptness of the learning curve in the Explicit-Promoting condition (though it did not significantly shorten response times the way it did in Experiment 1).

5 Experiment 3

Experiments 1 and 2 used a scenario in which the phonotactic pattern distinguished between within-language categories. In Experiment 3 we turn to a different function of phonotactics, distinguishing well-formed words of a language from ill-formed words which are not possible in the language. Rather than train participants outright to classify words as well- or ill-formed, Experiment 3 framed the training task as the ecologically more-natural one of vocabulary learning.

Signatures of explicit learning (Table 1), though reduced, were still found in the Implicit-Promoting conditions of Experiments 1 and 2. A different paradigm that encourages incidental category learning might reduce or abolish rule-seeking, rule-stating, or rule correctness. To achieve this, Experiment 3 uses a vocabulary-learning task to direct attention away from general category properties and towards individual category members (Love, 2002; Wattenmaker, 1991). The explicit system, if used at all, is thus preoccupied with individual word-meaning pairs and has less capacity for formulating and testing hypotheses about the well-formedness of the words. If the high rate of explicit learning in the Implicit-Promoting conditions of Experiments 1 and 2 was due to intentional learning encouraged by the gender-learning paradigm, Experiment 3 should reduce or abolish rule-seeking, rule-stating, or rule correctness.

The procedure was similar to that of Experiment 1. On each trial in the Explicit-Promoting training condition, participants were shown a picture with two buttons below it. When moused over, one button played the pattern-conforming word associated with the picture, while the other played a non-conforming foil. Correct-incorrect feedback was provided as in previous experiments. On each trial in the Implicit-Promoting condition, one picture was presented, with a single button. Mousing over the button played the (pattern-conforming) name for the picture. The critical properties were the three which elicited the highest rates of Correct or Approximate Stating among Staters in Experiment 1 (Table 8): two vs. three syllables, initial vs. non-initial stress, and stops vs. fricatives.

Participants in the Explicit-Promoting condition were told that the correct names sounded systematically different from the incorrect names. Those in the Implicit-Promoting condition, however, were told only that they would be learning the names for objects. Since learning in the Implicit-Promoting condition was focused on the word-picture associations, the phonotactic pattern would be acquired through incidental learning if at all. The test phase was identical to that of Experiment 1, except that participants were instructed to

choose their best guess as to the correct name for each word, based on which option sounded more like a word of the language they had been learning.

A total of 96 people participated via Mechanical Turk. Data from 15 participants was excluded (6 reported that they deliberately chose test items to sound different from training, 4 that they took written notes, 2 that their first language was not English, and 3 reported more than one of these), leaving 81 valid participants (43 in the Explicit-Promoting condition and 38 in the Implicit-Promoting condition).

5.1 Results

5.1.1 Hypothesis 1: Effect of training condition on rule-seeking and -stating

Both rule-seeking and rule-stating persisted in the Implicit-Promoting condition (Figure 6 and Table 23). Participants in the Implicit-Promoting condition were numerically less likely than those in the Explicit-Promoting condition to be Rule-Seekers, but the difference was not even marginally significant by Fisher's exact test (two-sided, $p = 0.1498$). Implicit-Promoting participants were, however, significantly less likely than Explicit-Promoting participants to be Rule-Staters (Fisher's exact test, two-sided, $p = 0.03438$).

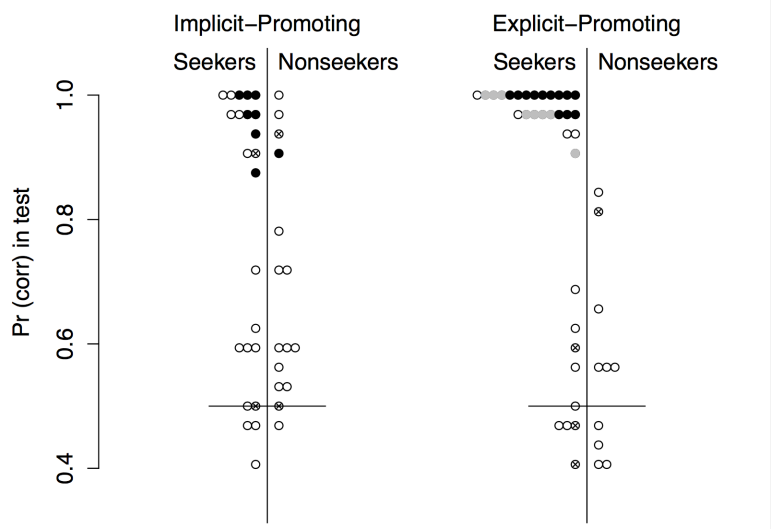


Figure 6: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 3. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

5.1.2 Hypothesis 2: Effect of rule-seeking on rule-stating and rule correctness

In the Explicit-Promoting condition, Seekers were significantly more likely than Non-Seekers to be Staters, as shown by the large and very significant effect of *Seeker* (Table 25). The two model coefficients expressing

	Rule-Seeker		Rule-Stater	
Training Condition	T	F	T	F
Explicit-Promoting	33	10	24	19
Implicit-Promoting	23	15	13	25

Table 23: Rule-Seeking and Rule-Stating as a function of Training Condition, Experiment 3

	Explicit-Promoting		Implicit-Promoting	
	Seekers	Non-Seekers	Seekers	Non-Seekers
Non-Staters	10	9	13	12
Staters	23	1	10	3
Correct Staters	12	0	7	1
Approximate Staters	8	0	0	0
Incorrect Staters	3	1	3	2

Table 24: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 3

differences between the training conditions, *Implicit-Promoting* and the two-way interaction, were both non-significant, indicating that the Implicit-Promoting condition did not differ significantly from the Explicit-Promoting condition. The same non-difference was found for *Correct or Approximate Stater* as the dependent variable (Table 26).

5.1.3 Hypothesis 3: Effect of rule correctness on generalization

Generalization performance by Non-Staters and Incorrect Staters was significantly above chance in both the Explicit-Promoting and Implicit-Promoting conditions (Table 27). As in the two previous experiments, *Rule Correctness* had a large and significant facilitating effect on correct (pattern-conforming) responses on the generalization test. Unlike in previous experiments, this effect was significantly reduced, though not eliminated, in the Implicit-Promoting condition. Figure 6 shows that the bulk of the difference between conditions falls at the top of the range: The Implicit-Promoting condition has fewer perfect performers and more between 90% and 98% correct. The basic fact found in Experiments 1 and 2 persists in the vocabulary-learning paradigm: The more correct the stated rule, the more correct the test-phase responses.

Coefficient	Estimate	Std. Error	χ^2 value	$\Pr(>\|z\|)$	
(Intercept)	-1.8458	0.9214	6.4864	0.0108	*
Seeker	2.6514	0.9955	11.0291	0.0009	***
Implicit-Promoting	0.5728	1.1132	0.3019	0.5826	
Seeker × Implicit-Promoting	-1.6298	1.2481	1.9628	0.1612	

Table 25: Fitted Firth logistic-regression model for Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 3

Coefficient	Estimate	Std. Error	χ^2 value	Pr($>|z|$)	
(Intercept)	-3.0445	1.5181	11.1812	0.000826	***
Seeker	3.4622	1.5592	12.4326	0.000421	***
Implicit-Promoting	0.7758	1.7576	0.2347	0.628049	
Seeker × Implicit-Promoting	-1.9820	1.8488	1.4531	0.228022	

Table 26: Fitted Firth logistic-regression model for Correct and Approximately-Correct Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 3

Coefficient	Estimate	Std. Error	t value	Pr($>|t|$)	
(Intercept)	0.5155	0.1717	3.002	0.003611	**
Rule Correctness	5.2520	0.6526	8.048	8.23e-12	***
Implicit-Promoting	0.2876	0.2404	1.196	0.235315	
Implicit-Promoting × Rule Correctness	−2.9517	0.7741	-3.813	0.000275	***

Table 27: Summary of fixed effects for the logistic-regression model for pattern-conformity of generalization-test responses, Experiment 3 (2592 responses from 81 participants).

5.1.4 Hypotheses 4 and 5: Effect of correct rule-stating on abruptness and response time

Although the Explicit-Promoting condition in this experiment was very similar to those in the previous two experiments, the negative effect of *Rule Correctness* on training-phase performance was non-significant (Table 28). The acceleration of response times across the last error, found in Experiment 1, also failed to replicate (Table 29).

Coefficient	Estimate	Std. Error	z value	t value	Pr $>|t|$
(Intercept)	0.8685	0.3124	2.780	0.01	**
Rule Correctness	−0.2398	0.5149	−0.466	0.6453	

Table 28: Summary of the the 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 Type IV condition of Experiment 3. (344 responses from 28 participants, excluding one participant dropped for making no errors after the first trial.)

Coefficient	Estimate	Std. Error	t value	Pr $>|t|$	
(Intercept)	2.3453	0.1383	16.959	7.29e-15	***
Preceding	0.0273	0.0289	0.946	0.3536	
Rule Correctness	-0.0877	0.0853	-1.028	0.3143	
Preceding × Rule Correctness	-0.0370	0.0854	-0.433	0.6689	
log(Trial Number - 1)	-0.0714	0.0299	-2.382	0.0255	*

Table 29: Summary of the general linear model for log response time, correct responses from Solvers in the Explicit-Promoting condition of Experiment 3within 16 trials of their last error. (686 observations from 29 participants).

5.1.5 Discussion

If the gender-based Implicit-Promoting conditions of Experiments 1 and 2 had (contrary to the experimenters' intentions) favored intentional learning, and therefore explicit learning, then replacing gender learning with vocabulary learning should have amplified the differences between the Explicit- and Implicit-Promoting conditions of Experiment 3, and should have reduced or eliminated rule-seeking and rule-stating in the Implicit-Promoting condition of Experiment 3 compared to those of Experiments 1 and 2.

Rule-seeking and rule-stating were not eliminated or even reduced by the change in paradigm. The proportions of Seekers in the Implicit-Promoting conditions of Experiments 1, 2, and 3 were respectively 55%, 68%, and 61%, and the proportions of Staters were 36%, 23%, and 34%. On both these dimensions, Experiment 3 fell in between Experiments 1 and 2, and none of the differences even approached significance by Fisher's exact test. However, the effect of rule correctness on generalization was significantly attenuated in the Implicit-Promoting condition of Experiment 3 relative to the Explicit-Promoting condition. This effect (a negative coefficient for *Rule correctness*×*Training Condition*) was apparent in the first two experiments as well, but did not reach significance. We can conclude that the high rate of explicit learning in the Implicit-Promoting conditions of Experiments 1 and 2 was not due to the use of the gender-learning task, but to participants' inclinations.

6 Experiment 4

The Implicit-Promoting condition in Experiments 1–3 exemplified the the most-common familiarization condition in phonotactic learning experiments, unreinforced exposure to pattern-conforming instances. The Explicit-Promoting condition in those experiments was designed to be maximally different from the Implicit-Promoting condition. Experiment 4 replaces that Explicit-Promoting condition with the most-widely-used training condition in non-linguistic category-learning experiments, single-interval binary classification with feedback (e.g., Kurtz et al. 2013; Nosofsky, Gluck, Palmeri, McKinley, and Gauthier 1994; Shepard et al. 1961). The two training conditions thus compare a typical phonotactic learning experiment with the phonotactic analogue of a typical non-linguistic category-learning experiment.

6.1 Methods

The stimuli were the same as used in the previous experiments. Participants were recruited in the same way, and participants in previous experiments could not participate in Experiment 4. Each participant's critical feature was chosen from one of two vs. three syllables, first- vs. second-syllable stress, and stop

vs. fricative consonants. In other respects the artificial-"language" generation procedure was as in the previous experiments. Participants in both training conditions were instructed that the experiment involved grammatical gender in an artificial language.

The Implicit-Promoting training condition was identical to that of Experiment 1. The Explicit-Promoting training condition differed from those of previous experiments in this paper. The previous Explicit-Promoting conditions used a two-interval forced-choice task, whereas this one used single-interval classification. Explicit-Promoting participants were instructed that they would be learning to tell whether a word belonged to the target gender, and that if they found the pattern, they could get it right every time. Of the 32 training items, half were feminine and half were masculine; i.e., half pattern-conforming and half nonconforming. On each trial, the picture was displayed above two buttons, labelled "Yes" and "No". Mousing over either one played the same word. The task was to decide whether the word belonged to the target gender. Feedback was given. The 32 training trials were repeated up to 4 times, but training finished early if a participant reached the criterion of 4 consecutive correct 4-trial blocks.

The task in the test phase was different from both of the training-phase tasks. A pattern-conforming word-picture pair was presented alongside a non-conforming word-picture pair, and participants were instructed to listen to both words and choose the one they thought was more likely to belong to the target gender.

Questionnaires were scored as in Experiment 1. Answers to the new free-response Question 7 (Table 6), about "Aha!" moments, were combined with those from Questions 2 and 4 when scoring. As noted above in Section 3.1.3, this was done because material pertaining to one question was often found in the answer box for a different question.

Participants in the Implicit-Promoting condition thus experienced training and testing similar to those used by Richtsmeier (2011), where participants were familiarized with "Martian" animals and their names as picture-word pairs, and in the test rated novel word-picture pairs for similarity to the familiarization stimuli.

A total of 94 participants completed the experiment. Of these, 12 were excluded from analysis (3 reported a non-English L1, 7 reported taking written notes, 2 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 none were excluded for two or more of these reasons), leaving 82 valid participants.

6.2 Results

6.2.1 Hypothesis 1: Effect of training condition on rule-seeking and -stating

Results from all participants are plotted in Figure 7. Participants in the Implicit-Promoting condition were numerically less likely than those in the Explicit-Promoting condition to report rule-seeking and rule-stating, but neither difference even approached significance by Fisher's exact test (two-sided, $p = 1$ (sic) and $p = 0.647$, respectively).

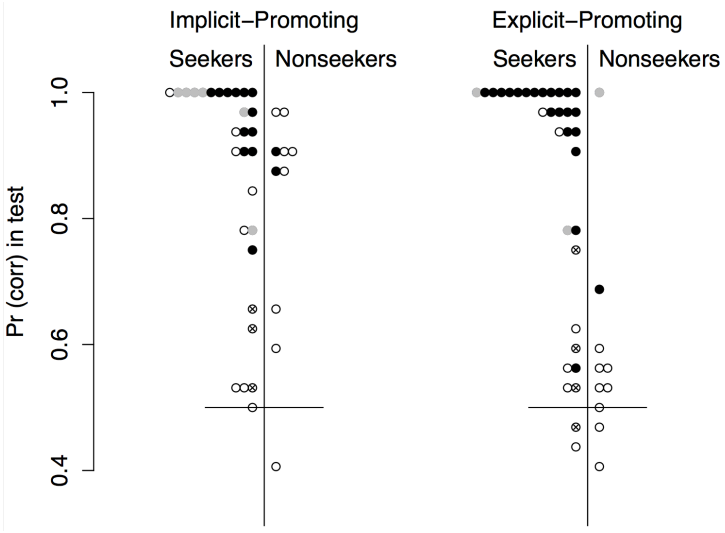


Figure 7: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 4. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

	Rule-Seeker		Rule-Stater	
Training Condition	T	F	T	F
Explicit-Promoting	33	10	29	14
Implicit-Promoting	29	10	24	15

Table 30: Rule-Seeking and Rule-Stating as a function of Training Condition, Experiment 4

6.2.2 Hypothesis 2: Effect of rule-seeking on rule-stating and rule correctness

In both training conditions, Seekers were significantly more likely than Non-Seekers to be Staters (Tables 31 and 32. As in Experiment 1and Experiment 5, most Staters were Correct Staters in both conditions. Like in Experiment 1, but unlike in Experiment 5, Explicit-Promoting training did not make Seekers more likely to be Staters (Tables 31 and 33).

	Explicit-Promoting		Implicit-Promoting	
	Seekers	Non-Seekers	Seekers	Non-Seekers
Non-Staters	6	8	7	8
Staters	27	2	22	2
Correct Staters	21	1	12	2
Approximate Staters	2	1	6	0
Incorrect Staters	4	0	4	0

Table 31: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 4

Coefficient	Estimate	Std. Error	χ^2 value	Pr($>\|z\|$)	
(Intercept)	-1.224	0.7545	3.45812	0.062941	.
Seeker	2.666	0.8748	12.3163	0.000449	**
Implicit-Promoting	1.499476e-15	1.0671	1.421085e-14	0.999999	
Seeker × Implicit-Promoting	-0.3437	1.2323	0.08416555	0.771729	

Table 32: Fitted logistic-regression model for Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 4

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	-1.223776	0.754	3.458120	0.062	.
Seeker	2.029401	0.843	7.464190	0.006	**
Implicit-Promoting	-9.095418e-16	1.067	1.421085e-14	0.999	
Seeker × Implicit-Promoting	-0.330202	1.194	0.082897	0.773	

Table 33: Fitted logistic-regression model for Correct and Approximate Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 4

6.2.3 Hypothesis 3: Effect of rule correctness on generalization

A logistic-regression model for pattern-conformity of generalization-test responses was fit using the same procedure as in Experiment 1 and Experiment 5 (Table 34). As in both of those experiments, Incorrect Staters and Non-Staters in both training groups performed significantly above chance on the test. The two previous experiments found a non-significant advantage for Implicit-Promoting over Explicit-Promoting Incorrect Staters and Non-Staters; in Experiment 4, it was significant. Correct and Approximate Staters once again did significantly better than Incorrect Staters and Non-Staters, with no significant difference between performance in the Explicit- and Implicit-Promoting conditions.

Coefficient	Estimate	Std. Error	t value	p	
(Intercept)	0.3822	0.1464	2.610	0.01085	*
Implicit-Promoting	0.7379	0.2721	2.712	0.00822	**
Rule Correctness	2.4920	0.4807	5.184	$1.66e-06$	***
Implicit-Promoting × Rule Correctness	−0.5254	0.6852	−0.767	0.44550	

Table 34: Summary of fixed effects for the logistic-regression model for pattern-conformity of generalization-test responses, Experiment 4. (2624 responses from 82 participants.)

6.2.4 Hypotheses 4 and 5: Effect of correct rule-stating on abruptness and response time

This time, no difference in abruptness was apparent between the Correct Staters and the others. The logistic-regression model finds no significant difference (Table 35), and the sign of the interaction term is in any case positive, the opposite of the prediction.

Coefficient	Estimate	Std. Error	t value	Pr $> \|t\|$	
(Intercept)	0.7336	0.3289	2.230	0.0349	*
Correct Stater	0.3498	0.3996	0.875	0.3897	

Table 35: Summary of the 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 4. (360 responses from 27 participants, excluding 1 more participant who made no errors after the first trial.)

The response-time acceleration for Correct Staters at the last error that we observed in Experiments 1 and 2 failed to replicate here. Participants in all conditions got faster as the experiment went on, but there were no significant differences between conditions (Table 36). This is not surprising: In Experiments 1 and 2, a participant who discovered a correct rule could take a shortcut, and respond after listening to only one of the two stimuli. In Experiment 4, where only one stimulus was presented per trial, no such shortcut was possible.

Coefficient	Estimate	Std. Error	t value	$\Pr > \|t\|$	
(Intercept)	2.1860455	0.1035459	21.112	$<$ 2e-166	***
Correctness	0.0008836	0.0194216	0.045	0.96411	
Preceding	0.0318113	0.0545997	0.583	0.56581	
Preceding $\times$ Correctness	0.0210398	0.0475628	0.442	0.66236	
log(Trial Number - 1)	-0.0809270	0.0271212	-2.984	0.00664	**

Table 36: Summary of fixed-effects portion of the linear mixed model for log response time, correct responses from Solvers in the Explicit-Promoting condition within 16 trials of their last error. (704 observations from 28 participants).

6.3 Discussion

In Experiment 4, the Implicit- and Explicit-Promoting conditions did not differ significantly in their effects on rule-seeking and rule-stating. As in the previous experiments, in both conditions rule-seeking facilitated rule-stating and rule correctness, and rule correctness facilitated generalization. Training which followed the standard phonotactic-learning paradigm and training which followed the standard non-linguistic category-learning paradigm thus had indistinguishable effects.

7 Experiment 5

In Experiments 1–3, the Implicit- and Explicit-Promoting condition differed in multiple ways: instruction, feedback, and whether each trial presented one conforming stimulus or a conforming-nonconforming pair. When Experiment 4 used a single word-picture pair per training trial in both conditions, the two conditions no longer differed in the rates of rule-seeking or rule-stating. That raises the possibility that what actually made the difference in Experiments 1-3 was not the instructions or the feedback, but the opportunity to compare conforming with non-conforming stimuli side by side. To test this hypothesis, Experiment 5 used the same feedback and instructions in both conditions, presented conforming-conforming pairs in the Implicit-Promoting condition, and presented conforming-nonconforming pairs in the Explicit-Promoting condition. Thus, both conditions provided feedback, but only the Explicit-Promoting condition allowed participants to compare a conforming with a non-conforming stimulus on each trial.

7.1 Methods

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.[9]

The critical feature was chosen from two/three syllables, first-/second-syllable stress, and stops/fricatives. Unlike in Experiments 1–4, 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 5 used two-alternative choice trials with feedback. On each training trial, a positive word-picture pair was matched with a negative word-picture pair. The participant saw 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). Each time all 32 positive and all 32 negative pairs had been presented, the positive pairs were randomly re-matched with negative pairs for the next cycle (thereby changing, on average, all but one matching, Zager and 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".

Questionnaires were scored as in Experiments 1–4. Answers to the new free-response Question 7 (Table 6), about "Aha!" moments, were combined with those from Questions 2 and 4 when scoring. As noted above in Section 3.1.3, this was done because material pertaining to one question was often found in the answer box for a different question.

7.2 Results

7.2.1 Hypothesis 1: Effect of training condition on rule-seeking and -stating

As in Experiment 1, Rule-Seekers and Rule-Staters were found in both training conditions (Figure 8). 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). As in Experiments 1–4, participants in the Explicit-Promoting condition were significantly more likely to be Rule-Staters ($p = 0.0006625$ respectively by Fisher's exact test, two-sided).

[9] In the interests of statistical power and expositional brevity, the analysis here combines data from two temporally-separate runs of the same identical experiment, an initial batch with 142 participants (109 valid), plus a subsequent batch of 87 participants (67 valid) that was run alongside Experiment 9 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 in the initial batch, drops to marginal significance when the two batches are analyzed together.

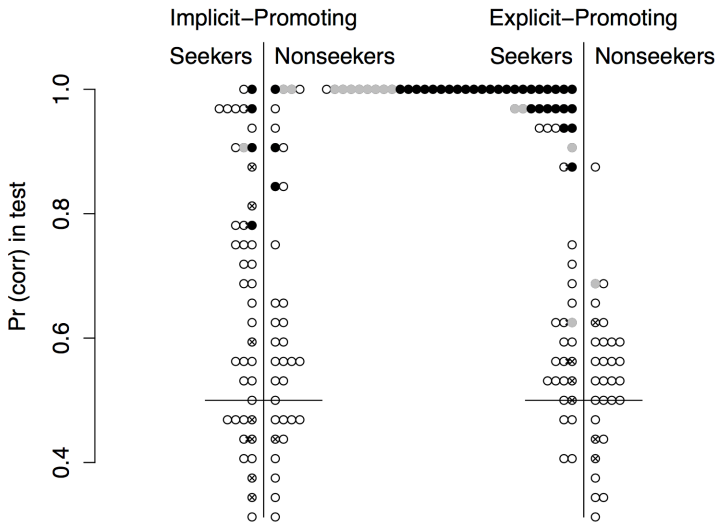


Figure 8: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 5. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

	Rule-Seeker		Rule-Stater	
Training Condition	T	F	T	F
Explicit-Promoting	69	30	53	46
Implicit-Promoting	43	34	21	56

Table 37: Rule-Seeking and Rule-Stating as a function of Training Condition, Experiment 5

7.2.2 Hypothesis 2: Effect of rule-seeking on rule-stating and rule correctness

In the Explicit-Promoting condition, Seekers were significantly more likely than Non-Seekers to be Staters (Tables 38 and 39), as in Experiments 1–4. However, the facilitating effect of rule-seeking on rule-stating was significantly reduced in the Implicit-Promoting condition.

	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 38: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 5

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	-1.773	0.5182	17.293	3.203166e-05	***
Seeker	2.654	0.5818	29.191	6.555829e-08	***
Implicit-Promoting	0.294	0.6805	0.195	6.581021e-01	
Seeker × Implicit-Promoting	-1.785	0.7968	5.251	2.192885e-02	*

Table 39: Fitted Firth logistic-regression model for Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 5

Table 38 also shows that Correct and Approximate Staters were found almost exclusively among Seekers in the Explicit-Promoting condition, and the logistic-regression model (Table 40) confirms a large and significant negative coefficient for the interaction of *Seeker* with *Implicit-Promoting*. The interaction term was large enough to entirely cancel out the effect of *Seeker* in the Implicit-Promoting condition.

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	-2.9789	0.8508	30.963	2.629486e-08	***
Seeker	3.4745	0.8862	33.933	5.704047e-09	***
Implicit-Promoting	1.2992	0.9726	2.187	1.391467e-01	
Seeker × Implicit-Promoting	-3.7408	1.1046	14.674	1.277706e-04	***

Table 40: Fitted Firth logistic-regression model for Correct and Approximate Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 5

7.2.3 Hypothesis 3: Effect of rule correctness on generalization

Table 41 shows that Incorrect Staters and Non-Staters performed above chance in both the Explicit- and Implicit-Promoting conditions. Correct and Approximate Staters did much better than Incorrect Staters

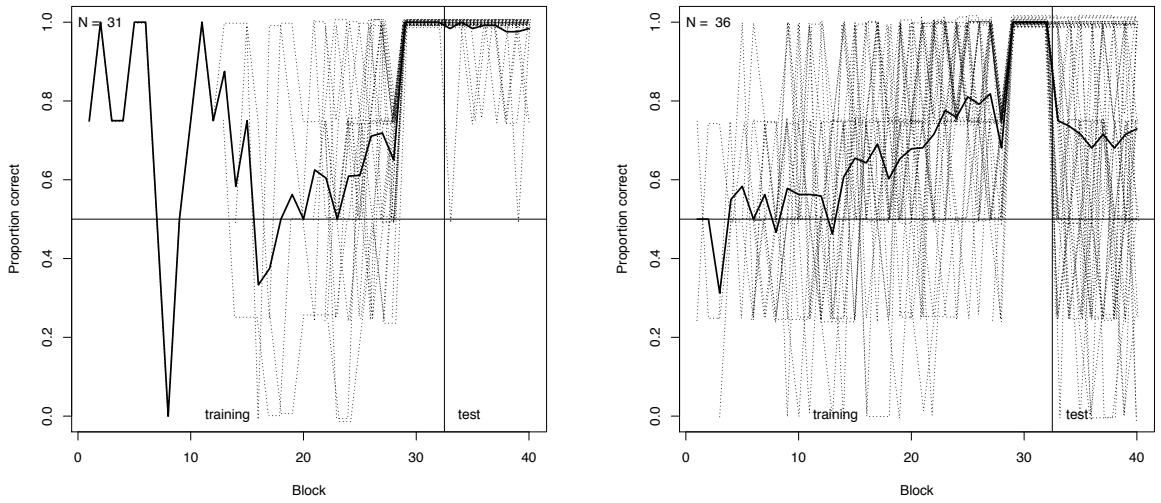
and Non-Staters in the Explicit-Promoting condition, but the benefit vanished in the Implicit-Promoting condition.

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 41: Summary of fixed effects for the logistic-regression model for pattern-conformity of generalization-test responses, Experiment 5 (5632 responses from 176 participants).

7.2.4 Hypotheses 4 and 5: Effect of correct rule-stating on abruptness and response time

Correct Staters show a more-abrupt performance jump across the last error, and their good performance persists throughout the test phase. Others (Non-Staters, Incorrect Staters, and Approximate Staters) show more-gradual improvement which tends to relapse in the test phase. The effect of Correct Stating on abruptness is confirmed statistically (Table 42). The response-time acceleration observed in Experiment 1 for Correct and Approximate Staters at the last error was replicated here. Even when the overall acceleration of responses over the course of the experiment was modelled out, there was a significant drop across the last error for Correct and Approximate Staters, but not for others (Table 43).



Correct Staters Others

Figure 9: Learning curves for Solvers in the Explicit-Promoting condition of Experiment 5, aligned to last error. Dashed lines are individuals, solid line is the mean across participants.

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

Table 42: Summary of the the 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 5. (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 > $\lvert t \rvert$	
(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 43: 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.)

7.3 Discussion

The overall pattern of results was almost unaltered from Experiment 1, except for two things. One is that the Explicit-Promoting condition no longer significantly facilitated rule-seeking compared to the Implicit-Promoting condition. The other is that Correct and Approximate Stating, which in Experiment 1 occurred frequently in both training conditions, was in Experiment 5 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.

8 Discussion: Experiments 1–5

The first five 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). The experiments, the main hypotheses tested, and the outcomes are summarized in Table 3.

Hypothesis 1: Rule-seeking and rule-stating are influenced (but not wholly determined) by instructions, feedback, and intention to learn. This hypothesis was supported by differences between the Explicit- and Implicit-Promoting conditions in Experiments 1–3, though not in Experiment 4. In all five experiments, both Rule-Seekers and Non-Rule-Seekers were found in both training conditions (Table 44). None of the 10 training conditions even came close to abolishing either approach. Seekers always formed a majority in every condition of every experiment. Relative to the Implicit-Promoting condition, the Explicit-Promoting

condition elicited higher rates of Seeking in Experiments 1 and 5, and higher rates of Stating in Experiments 1, 2, 3, and 5. Experiment 4 showed that varying rule instructions, feedback, and task was not always enough to affect the Seeking and Stating rates. Experiment 5 showed that when those factors are held constant, the Stating rate is higher, and the stated rules are more correct, when participants have the opportunity to compare a conforming and a non-conforming stimulus on each trial.

	Explicit-Promoting		Implicit-Promoting	
Experiment	Seekers:Nonseekers	Proportion Seekers	Seekers:Nonseekers	Proportion Seekers
1	54:9	0.86	41:33	0.55
2	21:5	0.81	15:7	0.68
3	33:10	0.76	23:15	0.61
4	33:10	0.77	29:10	0.74
5	69:30	0.70	43:34	0.56

Table 44: Ratio of Seekers to Non-Seekers and proportion of Seekers, Experiments 1–5

Hypothesis 2: Rule-stating is facilitated by rule-seeking. In all five experiments, self-report of rule-seeking was associated with a greater probability of reporting a rule, and of reporting an objectively correct or partly-correct rule. The straightforward interpretation is that subjective self-report of rule-seeking is accurate, and that participants were aware of whether they were or were not using a rule-seeking cognitive process. An alternative account of the facts is that rule-stating instead affects self-report of rule seeking: Unsuccessful rule-seekers may report not having sought a rule even though they did. We return to this alternative in Section 13.2 below, where evidence against it is presented.

Hypothesis 3: Stating a correct rule predicts better generalization performance. In all five experiments, Correct and Partly-Correct Staters gave significantly more pattern-conforming responses on the generalization test than did Non-Staters or Incorrect Staters. The effect was particularly dramatic among Staters who were also 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 generalization test, while that of the other Solvers fell sharply. 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 4: Correct rule-stating is associated with a more-abrupt learning curve. In Experiments 1, 2, and 5, 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. This finding excludes an alternative interpretation of the effect of correct rule-stating on generalization performance, namely, that participants' generalization responses were generated intuitively, and the stated rule was a retrospective account of their own behavior. The association between correct rule-stating and an abrupt performance jump during the

training phase shows that Correct Staters differ from others at an earlier point than predicted by this alternative. No significant difference in learning-curve abruptness was found in Experiments 3 or 4. That is somewhat puzzling in the case of Experiment 3, where the Explicit-Promoting condition was identical to that of Experiment 5; however, the numerical trend was in the predicted direction, and Experiment 3 had lower statistical power due to having fewer valid participants.

Hypothesis 5: Correct rule-stating is associated with response-time acceleration after the last error. This hypothesis was borne out in Experiments 1 and 5, but not in Experiments 2, 3, and 4. 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 and Rose (2007). Since two audio stimuli were presented on each training trial in Experiments 1–3, one positive and one negative, rule discovery could have allowed a participant to respond after listening to only one of them. That would also explain the non-replication in Experiment 4, where no time savings was possible because only one stimulus was presented per trial.

Together, the results of Experiments 1–5 support the two-process hypothesis reviewed in Section 2. The weakest link was in the effect of training condition: Training conditions in which participants were instructed to seek a rule and were given feedback indeed elicited greater rates of rule-seeking and rule-stating than those that did not (Experiments 1, 2, 3), but further probing to remove confounds showed that the factor which most strongly facilitated rule-seeking and rule-stating was the presentation of stimuli in positive-negative pairs, rather than participant instructions.

The other predictions of the two-process hypothesis were largely borne out. Rule-seeking and rule-stating occurred at non-negligible rates in every training condition of every experiment. Self-reported rule seeking was associated with significantly greater rates of rule-stating and more-correct stated rules in all five experiments. Correct rule-stating was associated with a more-abrupt learning curve (Experiments 1, 2, and 5), response acceleration after the last error (Experiments 1 and 5), and more correct generalization-test responses (all five experiments). In addition, participants who did not report rule-seeking and who did not state a rule nonetheless performed above chance on the generalization test, indicating that they had learned something about the pattern even if they did not (or could not) verbalize it. No significant experimental result directly contradicted any of the predictions.

9 Experiment 6

The implicit and explicit systems are hypothesized to have different architectures and hence different inductive biases, i.e., they are predicted to be good at learning different kinds of pattern (Section 2). Studies of

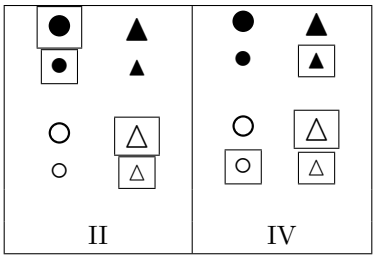


Figure 10: Examples of visual Type II and Type IV patterns.

visual learning have found that the abstract featural structure of a visual pattern can affect how hard it is to learn inductively, and that different patterns become harder or easier to learn when the experimental conditions favor implicit or explicit learning. Experiments Experiment 6–Experiment 9 ask whether the same is true of phonotactic learning.

A well-studied case is the contrast, illustrated in Figure 10, between the patterns on three binary features known for historical reasons as "Type II" and "Type IV" (Shepard et al., 1961). 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 an in-category prototype, e.g., "within one feature change of a small white triangle". When humans are asked to learn visual patterns of this sort via single-interval classification with feedback, the typical finding is that Type II patterns are easier than Type IV, in terms of trials to criterion and of total errors during training (Nosofsky, Gluck, et al., 1994; Shepard et al., 1961; J. D. Smith et al., 2004; Vigo, 2013).[10] Changing the experimental conditions so as to promote implicit learning reduces performance on 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 features. Since only two features are relevant for Type II, whereas three are relevant for Type IV, Type II is thus made easier to learn (Bradmetz & Mathy, 2008; Feldman, 2000, 2006; Kurtz et al., 2013; Lafond, Lacouture, & Mineau, 2007; Mathy & Bradmetz, 2004; Nosofsky, Palmeri, & McKinley, 1994; Shepard et al., 1961; Vigo, 2009). Feature-minimizing inductive biases have been independently proposed in phonology as a way of accounting for natural-language phenomena involving synchronic typology, diachronic change, and acquisition (particularly explicit examples include Chomsky and Halle 1968, 168, 221, 331, 334; Bach and Harms 1972; King 1969, 88–89; N. V. Smith 1973, 155–158; Gordon 2004; Hayes 1999; Hayes, Zuraw, Siptár, and Londe 2009; Kiparsky 1982; Pater and Staubs 2013; Pycha, Nowak, Shin, and Shosted

[10] *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 and Schwanenflugel (1981); Moreton et al. (2017).

2003), and more natural-language phonological classes can be stated as Type II patterns than as Type IV patterns (Moreton & Pertsova, 2014).

Both psychology and phonology thus give us theoretical reasons to expect Type II phonotactic patterns to be easier to learn than Type IV. However, in a phonotactic-learning study similar to the Implicit-Promoting condition of Experiment 1, the exact opposite was found: Type IV was significantly easier than Type II (Moreton et al., 2017, Experiment 1). This unexpected result was not due to a difference between how visual and phonological patterns are learned, because the result was replicated using visual analogues of the phonological stimuli (Moreton et al., 2017, Experiment 2). An obvious hypothesis to explain this unexpected result is that many participants in both experiments were relying on implicit learning. If so, then subdividing the participants into Rule-Seekers vs. Non-Seekers should reveal that Seekers do better on Type II than Type IV, at least relative to Non-Seekers.

Experiment 6 is similar to Experiment 1 and Experiment 2 except that, instead of all patterns being Type I, each participant receives either a Type II or a Type IV pattern. If the two systems used in phonological pattern learning function like those used in non-linguistic learning, then 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 6*).

9.1 Methods

The critical features were chosen from among two/three syllables, stops/fricatives, and labials/alveolars. 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, 44 in each of the Type II and Type IV conditions.

9.2 Results

Results from Experiment 6 were analyzed following the same procedures as for the corresponding analyses in the preceding experiments. Since no significant results were found in the analyses of Hypothesis 4 and Hypothesis 5 in this or in any subsequent experiment, the corresponding sections are omitted.

9.2.1 Hypothesis 1: Effect of training condition on rule-seeking and -stating

As in the preceding experiments using Type I patterns, rule-seeking and rule-stating occurred in both training conditions (Figure 11, Table 45). A Firth-penalized logistic-regression model with *Seeker* as the dependent variable and *Training Condition* and *Type* as predictors found no significant effect of either predictor and no interaction (Table 46). However, in the Type II condition, the Implicit-Promoting group were significantly less likely than the Explicit-Promoting group to be Rule-Staters. No significant effects of or interactions with Type were found (Table 47).

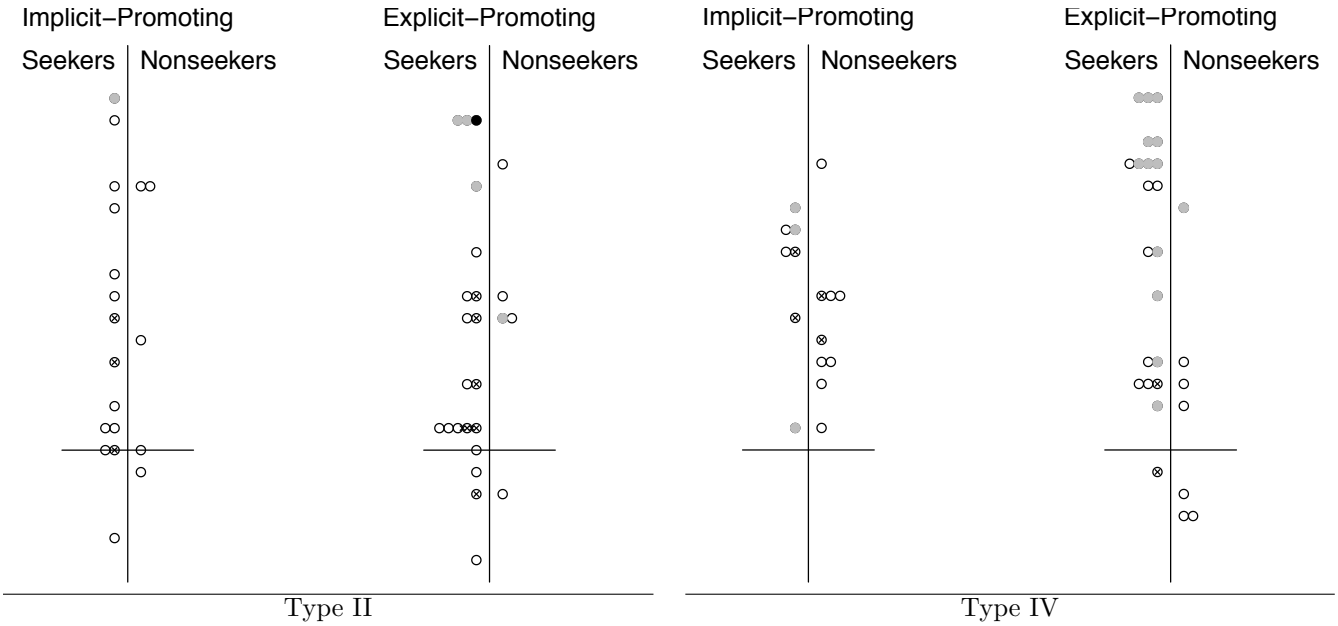


Figure 11: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 6. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

		Rule-Seeker		Rule-Stater	
Training Condition	Type	T	F	T	F
Explicit-Promoting	II	20	5	13	12
	IV	21	7	15	13
Implicit-Promoting	II	14	5	4	15
	IV	7	9	7	9

Table 45: Rule-Seeking and Rule-Stating as a function of Training Condition, Experiment 6

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 46: Fitted Firth logistic-regression model for Rule-Seeking as a function of Training Condition and Type, Experiment 6

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 47: Fitted Firth logistic-regression model for Rule-Stating as a function of Training Condition and Type, Experiment 6

9.2.2 Hypothesis 2: Effect of rule-seeking on rule-stating and rule correctness

Figure 11 Table 48 shows that 70% of the Seekers in the Explicit-Promoting condition (36/51) were Staters, as opposed to 27% (13/47) of other participants (non-Seekers in both training conditions, plus Seekers in the Implicit-Promoting condition). A preliminary logistic-regression model, analogous to the one in Table 12, was fit to the data, with *Stater* as the dependent variable and *Seeker*, *Training Condition*, and *Type* and their interactions as predictors. The analysis found no effect of, nor interaction with, *Type*. That predictor was therefore dropped and the model was refit to yield Table 49. Seekers were again significantly more likely than Non-Seekers to be Staters, but no significant effect of, nor interaction with, Training Condition was found.[11]

	Explicit-Promoting						Implicit-Promoting					
	Seekers			Non-Seekers			Seekers			Non-Seekers		
	II	IV	All	II	IV	All	II	IV	All	II	IV	All
Non-Staters	8	7	15	4	6	10	10	2	12	5	7	12
Staters	12	14	36	1	1	2	4	5	9	0	2	2
Correct Staters	1	0	1	0	0	0	0	0	0	0	0	0
Approximate Staters	3	12	15	1	1	2	1	3	4	0	0	0
Incorrect Staters	8	2	10	0	0	1	3	2	5	0	2	2

Table 48: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 6

[11] Correct and Approximate Stating occurred almost entirely among Seekers (22) rather than Non-Seekers (2). Since there were no Correct Staters among Non-Seekers in either training condition, and no Approximate Staters among Non-Seekers in the Implicit-Promoting condition, a logistic-regression model could not be fit.

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	-1.4350	0.7324	5.2934	0.0072	**
Seeker	1.9713	0.8007	8.0445	0.0003	***
Implicit-Promoting	-0.1743	1.0250	0.0311	0.7701	
Seeker × Implicit-Promoting	-0.6363	1.1617	0.3195	0.3047	

Table 49: Fitted Firth logistic-regression model for Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 6

9.2.3 Hypothesis 3: Effect of rule correctness on generalization

The mode of perfect generalization performance among Correct Staters that was seen in Experiments 1–5 is conspicuously absent here (Figure 11). There were so few Correct and Approximate Staters in the Implicit-Promoting condition, particularly in Type II, that a logistic-regression model with *Rule Correctness* as a predictor could not be accurately fit. The analysis was therefore restricted to the Explicit-Promoting condition alone. Pattern-conforming responses were coded as 1, non-conforming responses as 0. Pattern type was dummy-coded, with Type II (expected to produce the fewest pattern-conforming responses) as the reference category. The fitted model is shown in Table 50. 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 in either of these respects, as shown by the non-significant coefficients for *IV* and its interaction with *Rule Correctness*. Thus, the effect found in the earlier experiments, that correct or approximate stating predicts more pattern-conforming responses on the generalization test, extends to Type II and Type IV patterns in the Explicit-Promoting condition.

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 50: Summary of fitted logistic-regression model for pattern-conformity of generalization-test responses, Experiment 6, Explicit-Promoting condition only. Type II is the reference category. (1696 responses from 53 participants.)

9.2.4 Hypothesis 6: Effect of rule-seeking on relative difficulty of patterns

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). The results from Experiments 1–4 show that participants in the same condition can differ widely in how they learn, and validate the use of self-reported rule-seeking as a more-sensitive individual-level index of explicit learning. Figure 11 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 benefitted the least. Among Seekers in both training conditions, performance on Type II is well below that on Type and IV. These observations are confirmed by a logistic-regression model (Table 51), in which the only significant terms are the intercept and the interactions *I* × *Seeker* and *IV* × *Seeker*.

Coefficient	Estimate	Std. Error	z value	Pr(> |z|)	
(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 51: Summary of fitted logistic-regression model for pattern-conformity of generalization-test responses, Experiment 6. Type II is the reference category. (2816 responses from 88 participants.)

9.3 Discussion

Some results from Experiments 1–4 were replicated with these more-complex target patterns. As in Experiments 2–5, the Explicit-Promoting condition did not significantly facilitate self-reported rule-seeking. Rule-seeking and rule-stating occurred in both training conditions and for both Type II and Type IV. The Explicit-Promoting condition and rule-seeking each facilitated rule-stating. The Correct Staters, who in the earlier experiments formed a mode at 100% in the distribution of test-phase performance, were absent from Experiment 6, presumably because the correct rules were harder to find or to state. The Approximately-Correct Staters did show better generalization performance than Non-Staters and Incorrect Staters, as before, but their learning curves were not significantly more abrupt, and their response times did not shorten significantly after the last error. The lack of abruptness and response-time effects could simply be because the

independent variable only ranged up to Approximate Staters. More interestingly, it might also be a sign that rules for Type II and Type IV are found incrementally rather than all at once.

Experiment 6 found that self-reported rule-seeking improves performance on Type IV so much that it exceeds performance on Type II. This is unexpected under models of rule-based learning which incorporate a bias towards patterns that depend on fewer features (see Section 2, above). A post-hoc explanation for the reversal will be discussed in Section 13.

10 Experiment 7

The results of Experiment 6 were surprising enough that Experiment 7 was done to see if they would replicate. In Experiment 6, some of the Type II patterns used the two features fricatives/stops and labial/coronal, which were both realized on the consonants. Type II patterns have previously been found to be significantly easier when both relevant features are realized in the same segment position than when they are realized on two different segment positions (Moreton et al., 2017, Exp. 1). That might have made those patterns easier relative to the Type IV problems, which always involved the syllable-length feature as well. Experiment 7 therefore used first- vs. second-syllable stress in place of Experiment 6's labial vs. coronal consonants.

10.1 Methods

The stimuli, instructions, and procedure were identical to those of Experiments 1, 2, and 7, except that each participant was randomly assigned a Type II or Type IV pattern. The same critical features were used as in Experiment 3 (two/three syllables, first-/second-syllable stress, stops/fricatives). 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.

10.2 Results

10.2.1 Hypothesis 1: Effect of training condition on rule-seeking and -stating

As in all previous experiments, rule-seeking and rule-stating occurred in both training conditions and in both pattern-type conditions (Figure 12, Table 52). A Firth-penalized logistic-regression model with *Seeker* as the dependent variable and *Training Condition* and *Type* as predictors found no significant effect of either predictor and no interaction (Table 53). 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 54). 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.

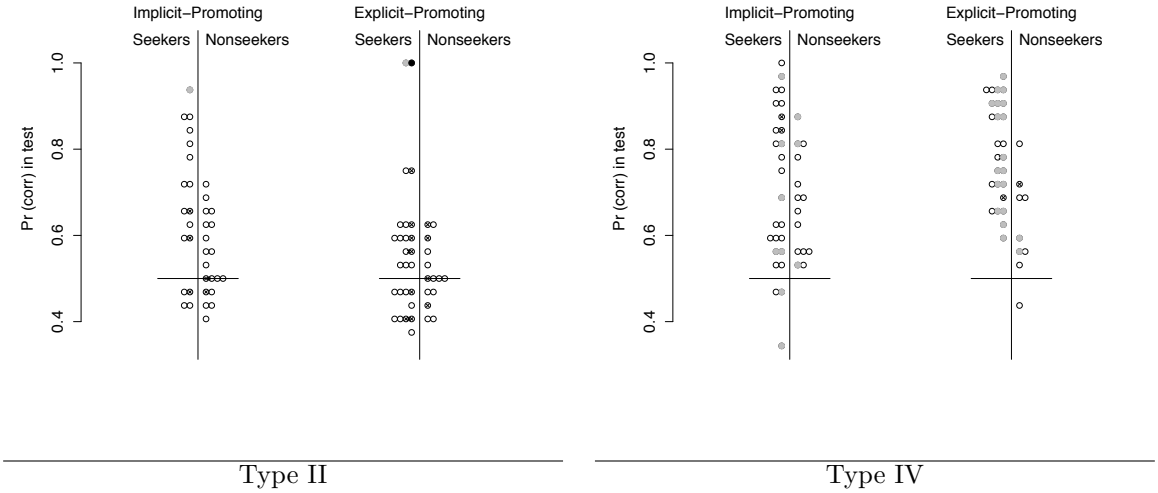


Figure 12: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 7. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

		Rule-Seeker		Rule-Stater	
Training Condition	Type	T	F	T	F
Explicit-Promoting	II	26	14	15	25
	IV	26	9	24	11
Implicit-Promoting	II	17	19	7	29
	IV	26	14	24	12

Table 52: Rule-Seeking and Rule-Stating as a function of Training Condition, Experiment 7.

Coefficient	Estimate	Std. Error	χ^2 value	p
(Intercept)	0.6029	0.3307	3.5641	0.0590
Implicit-Promoting	-0.7112	0.4698	2.3867	0.1223
IV	0.4228	0.5064	0.7244	0.3946
IV × Implicit-Promoting	0.2883	0.6908	0.1786	0.6725

Table 53: Fitted Firth logistic-regression model for Rule-Seeking as a function of Training Condition and Type, Experiment 7.

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 54: Fitted Firth logistic-regression model for Rule-Stating as a function of Training Condition and Type, Experiment 7.

10.2.2 Hypothesis 2: Effect of rule-seeking on rule-stating and rule correctness

Figure 12 Table 55 shows that Seekers were more likely than Non-Seekers to be Staters, regardless of Training Condition or Type. This observation is confirmed by the logistic-regression model shown in Table 56.

	Explicit-Promoting						Implicit-Promoting					
	Seekers			Non-Seekers			Seekers			Non-Seekers		
	II	IV	All	II	IV	All	II	IV	All	II	IV	All
Non-Staters	15	5	20	10	6	16	13	17	30	16	11	27
Staters	11	21	32	4	3	7	4	9	13	3	3	6
Correct Staters	1	0	1	0	0	0	0	0	0	0	0	0
Approximate Staters	1	17	18	0	2	2	1	7	8	0	3	3
Incorrect Staters	9	4	13	4	1	5	3	2	5	3	0	3

Table 55: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 7.

Coefficient	Estimate	Std. Error	χ^2 value	Pr($>\|z\|$)	
(Intercept)	−0.7884574	0.4498572	3.4589187	0.06291144	.
Seeker	1.2492726	0.5324012	6.0826904	0.01365127	*
Implicit-Promoting	−0.6539265	0.6311460	1.1180931	0.29032943	
Seeker × Implicit-Promoting	−0.6219257	0.7673141	0.6752501	0.41122714	

Table 56: Fitted Firth logistic-regression model for Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 7.

A preliminary analysis of Correct and Approximate Stating using *Training Condition*, *Type*, and *Seeker* as predictors found a significant effect of *Seeker* and no significant interactions of any variable with *Type*, so the model was simplified by omitting *Type* and refit (Table 57). It shows that Seekers were significantly more likely than Non-Seekers to be Correct or Approximate Staters, with no significant difference between Training Conditions.

10.2.3 Hypothesis 3: Effect of rule correctness on generalization

The mode at 100% pattern-conforming generalization responses which was found in Experiments 1–5, and which disappeared with the switch to more-complex pattern types in Experiment 6, was again absent here.

Coefficient	Estimate	Std. Error	χ^2 value	Pr($>|z|$)	
(Intercept)	-2.15176211	0.6825871	17.23221	0.000033	***
Implicit-Promoting	-0.01320161	0.8911105	0.000228	0.987961	
Seeker	1.61063113	0.7406876	6.363428	0.011649	*
Seeker × Implicit-Promoting	-0.87513395	1.0129057	0.784275	0.375836	

Table 57: Fitted Firth logistic-regression model for Correct and Approximate Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 7.

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 (Table 58). The intercept means that participants in the Explicit-Promoting Type IV condition tended to give pattern-conforming responses on the generalization test. No other term differed significantly from zero, indicating that neither rule correctness nor training condition had any measurable influence on test-phase performance. The ineffectiveness of Rule Correctness may be due in part to its small range; since there were no Correct Staters, Rule Correctness never exceeded 0.5.

Coefficient	Estimate	Std. Error	z value	Pr($>|z|$)	
(Intercept)	0.98710	0.17261	5.719	2.36e-07	***
Rule Correctness	0.49555	0.49117	1.009	0.316	
Implicit-Promoting	−0.08463	0.21692	−0.390	0.698	
Implicit-Promoting × Rule Correctness	−0.95159	0.77651	−1.225	0.224	

Table 58: Summary of fitted logistic-regression model for pattern-conformity of generalization-test responses Experiment 7, Type IV condition only. Incorrect or no rule in the Explicit-Promoting condition is the reference category. (2400 responses from 75 participants.)

10.2.4 Hypothesis 6: Effect of rule-seeking on relative difficulty of patterns

Cell means for test-phase performance are shown in the corresponding fitted logistic-regression model in Table 59. 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 $IV \times Seeker$). This much is consistent with what was found in Experiment 6. 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 $IV \times Seeker$).

10.3 Discussion

The outcome of Experiment 7 was very much like that of Experiment 6. In particular, the same novel effect seen in Experiment 6 is replicated in Experiment 7: In the Explicit-Promoting condition, self-reported

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 59: Summary of fitted logistic-regression model for pattern-conformity of generalization-test responses, Experiment 7. Type II is the reference category. (4832 responses from 151 participants.)

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 generalization performance. These results again directly contradict Hypothesis 6.

11 Experiment 8

Experiment 8 substituted the vocabulary-learning paradigm of Experiment 3 in place of the gender-learning paradigm. Of the 185 participants who finished, data from 52 was excluded (39 reported that they deliberately chose test items to sound different from training, 8 that they took written notes, 1 that their first language was not English, and 4 reported more than one of these), leaving 133 valid participants (75 Explicit-Promoting and 58 Implicit-Promoting).

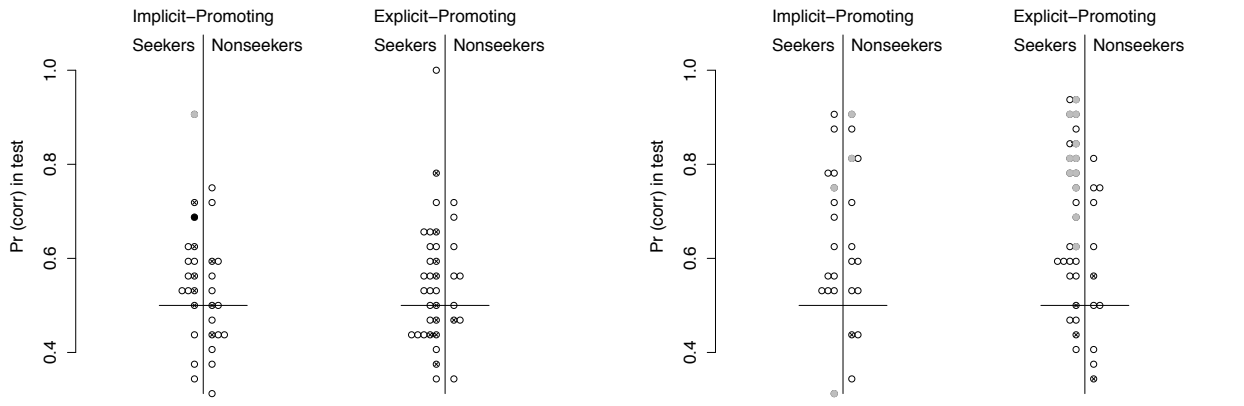
11.1 Results

11.1.1 Hypothesis 1: Effect of training condition on rule-seeking and -stating

As in all previous experiments, both rule-seeking and rule stating were found in both training conditions (Figure 13, Table 60). Among participants in the Type II condition, rule-seeking was significantly less common with Implicit-Promoting than Explicit-Promoting training (Table 60), and the Type IV condition did not differ significantly from the Type II condition. Rule-stating was not significantly affected by either factor (Table 62).

11.1.2 Hypothesis 2: Effect of rule-seeking on rule-stating and rule correctness

A preliminary logistic-regression model was fit to the data in Table 63, with *Stater* as the dependent variable and *Seeker*, *Training Condition*, and *Type* and their interactions as predictors. The analysis found no effect



Type II Type IV

Figure 13: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 8. 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%.

		Rule-Seeker		Rule-Stater	
Training Condition	Type	T	F	T	F
Explicit-Promoting	II	28	9	11	26
	IV	27	11	16	22
Implicit-Promoting	II	16	15	10	21
	IV	14	13	5	22

Table 60: Rule-Seeking and Rule-Stating as a function of Training Condition, Experiment 8.

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	1.0986	0.3796	9.9417	0.0016	**
Implicit-Promoting	-1.0360	0.5227	4.1782	0.0409	*
IV	-0.2267	0.5202	0.1956	0.6582	
IV × Implicit-Promoting	0.2357	0.7404	0.1045	0.7464	

Table 61: Fitted Firth logistic-regression model for Rule-Seeking as a function of Training Condition and Type, Experiment 8.

Coefficient	Estimate	Std. Error	χ^2 value	p
(Intercept)	-0.8347	0.3578	6.0852	0.0136
IV	0.5246	0.4856	1.2115	0.2710
Implicit-Promoting	0.1181	0.5238	0.0523	0.8190
IV × Implicit-Promoting	-1.2167	0.7853	2.5396	0.1110

Table 62: Summary of fitted logistic-regression model for rule-stating as a function of pattern type and training condition, Experiment 8.

of, nor interaction with, *Type*. That predictor was therefore dropped and the model was refit. Rule-Seeking significantly facilitated Rule-Stating in the Implicit-Promoting condition, and there were no significant effects of, nor interactions with, Type (Table 64).

	Explicit-Promoting						Implicit-Promoting					
	Seekers			Non-Seekers			Seekers			Non-Seekers		
	II	IV	All	II	IV	All	II	IV	All	II	IV	All
Non-Staters	18	13	31	8	9	17	9	12	21	12	10	22
Staters	10	14	24	1	2	3	7	2	9	3	3	6
Correct Staters	0	0	0	0	0	0	1	0	1	0	0	0
Approximate Staters	0	11	11	0	0	0	1	2	3	0	2	2
Incorrect Staters	10	3	13	1	2	3	5	0	5	3	1	4

Table 63: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 8.

Since only two Type II participants stated a Correct or Approximately Correct rule, the analysis of Correct and Approximate Stating as a function of Training Condition and Seeking was restricted to participants in the Type IV condition (Table 65). It shows that Seekers were significantly more likely than Non-Seekers to be Correct or Approximate Staters. There was a large but only marginally-significant reduction of the Seeker advantage in the Implicit-Promoting condition.

11.1.3 Hypothesis 3: Effect of rule correctness on generalization

Because the Type II condition had only two Correct or Approximate Staters, a logistic-regression model with *Rule Correctness* as a predictor could not be accurately fit. The analysis was therefore restricted to the Type IV condition alone (Table 66). The modest but significant coefficient for the intercept shows that Type IV Explicit-Promoting participants favored pattern-conforming test items, even though they did not verbalize a rule. Correct and Approximate Stating significantly increased the probability of pattern-conforming generalization responses by a large amount. No other significant effects or interactions were found.

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	-1.6094	0.6000	10.1886	0.0014	**
Seeker	1.3581	0.6586	5.2654	0.0217	***
Implicit-Promoting	0.3677	0.7519	0.2551	0.6134	
Seeker × Implicit-Promoting	-0.9331	0.8922	1.1691	0.2795	

Table 64: Fitted Firth logistic-regression model for Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 6.

Coefficient	Estimate	Std. Error	χ^2 value	Pr(> |z|)	
(Intercept)	-3.1354	1.5088	12.4786	0.00041	***
Seeker	2.7744	1.5587	6.7936	0.00914	**
Implicit-Promoting	1.6094	1.6736	1.3015	0.25393	
Seeker × Implicit-Promoting	-2.8578	1.8623	3.3167	0.06857	

Table 65: Fitted Firth logistic-regression model for Correct and Approximate Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 8, Type IV condition only.

Coefficient	Estimate	Std. Error	z value	Pr(> |z|)	
(Intercept)	0.3939	0.1256	3.128	0.00263	**
Rule Correctness	2.0349	0.4298	3.611	1.35e-05	***
Implicit-Promoting	0.1577	0.1844	0.499	0.39570	
Rule Correctness × Implicit-Promoting	−1.4880	1.1971	−0.478	0.21864	

Table 66: Summary of fitted logistic-regression model for pattern-conformity of generalization-test responses, Experiment 8, Type IV condition only. Explicit-Promoting is the reference category. (2080 responses from 65 participants.)

11.1.4 Hypothesis 6: Effect of rule-seeking on relative difficulty of patterns

The novel finding of Experiments 6 and 7, that rule-seeking facilitates Type IV relative to Type II instead of the other way around, was not replicated in Experiment 8. The fitted model is shown in Table 67. None of the coefficients differed significantly from zero. The sign of the critical term ($IV \times Seeker$) is positive, as predicted, but it is not significantly greater than zero ($p = 0.119$).

Coefficient	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.1950	0.1504	1.297	0.197
IV	0.1142	0.2468	0.463	0.644
Implicit-Promoting	-0.1617	0.1930	-0.838	0.404
Seeker	0.0110	0.1826	0.060	0.952
IV × Implicit-Promoting	0.3941	0.3441	1.145	0.254
IV × Seeker	0.4789	0.3048	1.571	0.119
Seeker × Implicit-Promoting	0.2387	0.2565	0.931	0.354
Seeker × Implicit-Promoting × IV	-0.6337	0.4498	-1.409	0.161

Table 67: Summary of fitted logistic-regression model for pattern-conformity of generalization-test responses, Experiment 8. Type II is the reference category. (4256 responses from 133 participants.)

11.1.5 Discussion

Participants in Experiment 8 replicated in a vocabulary-learning paradigm some of the principal effects found in previous experiments, including occurrence of rule-seeking and rule-stating in both training conditions, facilitation of rule-seeking in the Explicit-Promoting condition, facilitation of rule-stating and rule correctness by rule-seeking, and facilitation of generalization by rule correctness. There was no mode in generalization

performance at 100% (in fact, no one gave 100% pattern-conforming responses).

Although Experiment 8did not outright contradict Hypothesis 6 the way Experiments 6 and 7 did, the results certainly gave no support for the hypothesis, and the nonsignificant numerical trend went in the wrong direction, i.e., towards rule-seeking improving performance on Type IV rather than on Type II.

12 Experiment 9

This experiment sought to replicate the rule-seeking effect on the Type IV advantage over Type II using the vocabulary-learning paradigm of Experiment 5.

12.1 Methods

The stimuli, instructions, and procedure were identical to those of Experiment 5, 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 people participated. 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 (55 Explicit-Promoting and 64 Implicit-Promoting).

12.2 Results

12.2.1 Hypothesis 1: Effect of training condition on rule-seeking and -stating

As in all previous experiments, both rule-seeking and rule stating were found in both training conditions (Figure 14, Table 68). However, as in Experiment 5, training condition did not affect either of these two variables significantly (Tables 69 and 70).

		Rule-Seeker		Rule-Stater	
Training Condition	Type	T	F	T	F
Explicit-Promoting	II	11	11	5	17
	IV	21	12	13	20
Implicit-Promoting	II	15	18	6	27
	IV	16	15	7	24

Table 68: Rule-Seeking and Rule-Stating as a function of Training Condition, Experiment 9

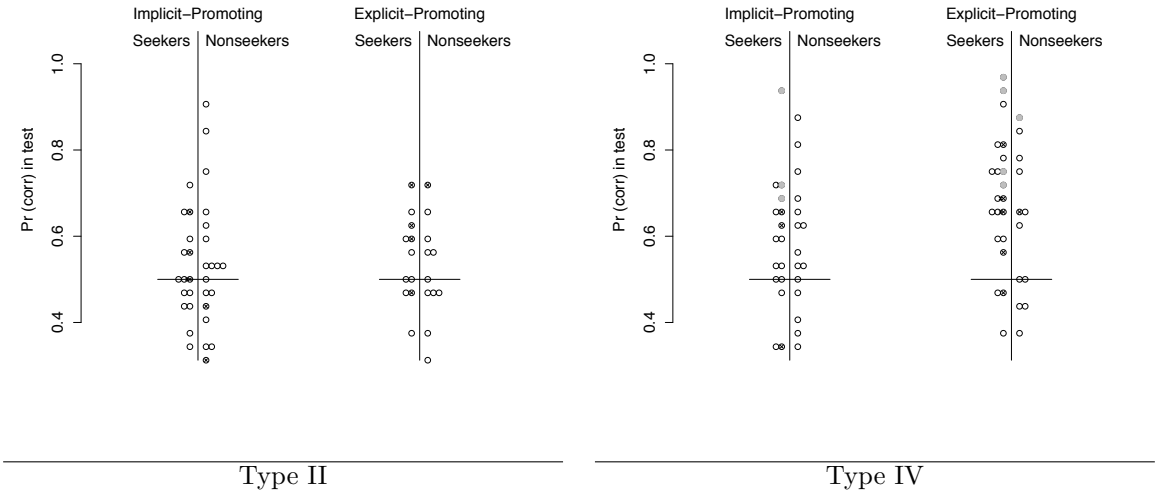


Figure 14: Test-phase performance as a function of training condition, rule-seeking, and rule-stating, Experiment 9. Plotting symbols: Black circle = Correct Stater, gray circle = Approximate Stater, crossed circle = Incorrect Stater, white circle = Non-Stater. A horizontal line segment marks the chance performance level of 50%.

Coefficient	Estimate	Std. Error	χ^2 value	p
(Intercept)	-2.27e-16	0.4264	0.0000	1.0000
Implicit-Promoting	-0.1769	0.5513	0.1070	0.7434
IV	0.5423	0.5587	0.9839	0.3212
IV × Implicit-Promoting	-0.3028	0.7506	0.1686	0.6812

Table 69: Fitted Firth logistic-regression model for Rule-Seeking as a function of Training Condition and Type, Experiment 9

Coefficient	Estimate	Std. Error	χ^2 value	p
(Intercept)	-1.1574	0.4998	6.5812	0.0103
Implicit-Promoting	-0.2849	0.6676	0.1881	0.6644
IV	0.7397	0.6135	1.5766	0.2092
IV × Implicit-Promoting	-0.4811	0.8672	0.3203	0.5714

Table 70: Fitted Firth logistic-regression model for Rule-Stating as a function of Training Condition and Type, Experiment 9

12.2.2 Hypothesis 2: Effect of rule-seeking on rule-stating and rule correctness

A preliminary analysis with *Rule-stater* as the dependent variable and *Training Condition* and *Type* as predictors found no significant interaction between the two predictors and no effect of *Type*, so the data in Table 68 was collapsed across the two Types and refit (Table 72). The large, negative, and significant intercept reflects the low rate of rule-stating among Non-Seekers in the Explicit-Promoting condition. The large, positive, and significant coefficient for *Seeker* shows that rule-seeking substantially increased the rate of rule-stating. No other coefficients were significant, indicating that the Implicit-Promoting condition did not differ significantly from the Explicit-Promoting condition. This again agrees with the results found in Experiment 5 with the same paradigm but a Type I pattern..

	Explicit-Promoting						Implicit-Promoting					
	Seekers			Non-Seekers			Seekers			Non-Seekers		
	II	IV	All	II	IV	All	II	IV	All	II	IV	All
Non-Staters	7	10	17	10	10	20	11	9	20	16	15	31
Staters	4	11	15	2	1	3	4	7	11	2	0	2
Correct Staters	0	0	0	0	0	0	0	0	0	0	0	0
Approximate Staters	0	4	4	0	1	1	0	3	3	0	0	0
Incorrect Staters	4	7	11	1	1	2	4	4	8	2	0	2

Table 71: Rule-Stating and correctness of stated rule as a function of Rule-Seeking, Experiment 9

Since there were no Correct or Approximate Staters in the Type II condition, the analysis of Correct and Approximate Stating as a function of Training Condition and Seeking was limited to participants in the Type IV condition. Aside from a large and significant negative intercept, indicating a very low rate of Correct and Approximate Stating, none of the terms in the model was significant (Table 73).

12.2.3 Hypothesis 3: Effect of rule correctness on generalization

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 logistic-regression model with the pattern-conformity of each generalization-test response as the dependent variable and *Rule*

Coefficient	Estimate	Std. Error	χ^2 value	p	
(Intercept)	-1.7676	0.5907	13.3312	0.0002	***
Seeker	1.6463	0.6888	6.9963	0.0081	**
Implicit-Promoting	-0.7660	0.8909	0.7792	0.3773	
Seeker × Implicit-Promoting	0.3093	1.0292	0.0940	0.7590	

Table 72: Fitted Firth logistic-regression model for Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 9

Coefficient	Estimate	Std. Error	χ^2 value	Pr(> |z|)	
(Intercept)	-2.0368	0.9035	8.7235	0.00314	**
Seeker	0.6787	1.0531	0.4796	0.48858	
Implicit-Promoting	-1.3971	1.7374	0.8071	0.36897	
Seeker × Implicit-Promoting	1.4053	1.9218	0.6338	0.42593	

Table 73: Fitted Firth logistic-regression model for Correct and Approximate Rule-Stating as a function of Rule-Seeking and Training Condition, Experiment 9, Type IV condition only

Correctness and *Training Condition* was fit as shown in Table 74. 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 were marginally less likely to give a pattern-conforming response, but there was no significant interaction between *Training Condition* and *Rule Correctness*, 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 74: Summary of fitted logistic-regression model for pattern-conformity of generalization-test responses, Experiment 9, Type IV only. (2048 responses from 64 participants.)

12.2.4 Hypothesis 6: Effect of rule-seeking on relative difficulty of patterns

Rule-seeking had no detectable effect on test-phase pattern-conformity of Type II vs. Type IV in either training condition, as shown in the statistical model in Table 75.

Coefficient	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	0.0682	0.1364	0.500	0.6183	
IV	0.4204	0.2418	1.738	0.0849	.
Implicit-Promoting	0.1058	0.2050	0.516	0.6068	
Seeker	0.1370	0.1788	0.766	0.4452	
IV × Implicit-Promoting	-0.2580	0.3285	-0.785	0.4339	
IV × Seeker	0.1976	0.3095	0.638	0.5245	
Seeker × Implicit-Promoting	-0.2360	0.2584	-0.914	0.3629	
Seeker × Implicit-Promoting × IV	-0.0798	0.4229	-0.189	0.8507	

Table 75: Summary of fitted logistic-regression model for pattern-conformity of generalization-test responses, Experiment 9. Type II is the reference category. (3808 responses from 119 participants.)

12.3 Discussion

Experiment 9, like Experiment 8, did not directly contradict Hypothesis 6 the way Experiments 6 and 7 did, but Experiment 9 did not support Hypothesis 6 at all. The nonsignificant numerical trend went in the "wrong" direction for Hypothesis 6, i.e., to the benefit of Seekers over Non-Seekers in the Type IV condition but not in the Type II condition. The main difference in outcomes between the gender-learning Experiments 7 and 6 on the one hand and the vocabulary-learning Experiments 8 and 9 on the other was that Seekers in both the Explicit- and Implicit-Promoting Type IV conditions — who in all four experiments outperformed Type IV Nonseekers on the generalization test — had worse generalization performance in the vocabulary-learning experiments, perhaps because a focus on memorizing vocabulary interfered with the search for an explicit rule.

13 Discussion: Experiments 6–9

Participants in the Type II/IV experiments (Experiments 6–9), like those in the Type I experiments (Experiments 1–5), 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 6– 8 relative to the Implicit-Promoting condition (Table 4). Rule-seeking invariably facilitated rule-stating and improved the correctness of stated rules, which in turn facilitated generalization performance in Experiments 6, 8, and 9. We have, therefore, ample reason to believe that in Experiments 6–9, as in Experiments 1–5, participants who reported rule-seeking tended as a group to use explicit learning more than those who did not.

Some findings of the Type I experiments were not replicated. Correct Staters were much rarer, reflecting the greater difficulty of the target pattern; the generalization 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 can be traced to the same source: Since the completely correct rule is harder to find and state explicitly, any effect of rule correctness is at best the weaker effect of an approximately-correct rule. The next two subsections elaborate on two aspects of this observation.

13.1 Type IV facilitates explicit search for relevant dimensions

The interaction between explicit learning and pattern structure was unexpected under existing theories of concept learning. The implicit system, which is typically modelled as learning by strengthening associations

between categories and cues or cue combinations, was expected to be more receptive to Type IV (family-resemblance), and less so to Type II (iff/xor), than the explicit system, which is typically modelled as learning by testing hypotheses in increasing order of featural complexity. Although that outcome has been found in non-linguistic concept learning (see Section 2), it was not found in any of these four experiments. In Experiments 8 and 9 using a vocabulary-learning paradigm, no significant difference was found (and the numerical trends went in the wrong direction). In Experiments 6 and 7 using a gender-learning paradigm, the exact opposite was found: Self-reported rule-seeking facilitated Type IV performance relative to Type II. How could that have happened?

We conjecture that explicit learners searched for the relevant dimensions ("attribute identification", Haygood and Bourne 1965) by serially testing one-dimensional rules. In the Type IV condition, this is a successful strategy: Each relevant dimension, individually, yields a one-feature rule that produces 75% correct responses during Explicit-Promoting training, and that characterizes 75% of the (all-positive) training items during Implicit-Promoting training. One-feature rules based on the irrelevant dimensions are 50% correct. A learner in the Type IV condition can use this information 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 describes 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 6, Type II, Explicit-Promoting condition)

Table 76 shows rates at which Seekers mentioned at least one of the pattern-relevant features in any free-response answer, regardless of whether they stated a rule or not. In all four experiments, these rates are much lower in the Type II condition than the Type IV condition. The most common type of Stater in Experiments 6–8 was an Approximate Stater in the Type IV condition and an Incorrect Stater in the Type II condition. Across all four II/IV experiments, there was a grand total of 3 Correct and 8 Approximate Type II Staters, versus 0 Correct and 68 Approximate Type IV Staters. The proportion of Correct Staters among Correct and Approximate Staters was significantly higher in the Type II than the Type IV condition ($p = 0.002$ by Fisher's Exact Test). It thus appears that participants in the Type IV condition readily found an approximate one-feature rule, but did not progress further, whereas those in the Type II condition had

great difficulty finding the relevant dimensions, but, once found, readily composed them into an exact rule.

	Experiment			
Type	6	7	8	9
II	0.03	0.07	0.05	0.00
IV	0.36	0.40	0.20	0.14

Table 76: Proportion of valid Seekers mentioning at least one pattern-relevant feature in any free-response question.

13.2 Rule-seeking predicts rule-stating

Hypothesis 2 states: "If the explicit system is indeed under voluntary control, then the products of that system (namely rules) ought to be reported more often by participants who report voluntary use of that system." We have interpreted the association between rule-seeking and rule-stating as corroborating this hypothesis. An alternative explanation, long familiar to observers of human nature, was suggested above (Section 8): Perhaps unsuccessful seekers report non-seeking. If this "sour grapes" alternative is true, then harder patterns ought to elicit lower rates of self-reported rule-seeking. In none of Experiments 6–9, however, was any significant difference found between the proportion of Seekers in the (harder) Type II and (easier) Type IV conditions.

As a more rigorous test, we included the even easier Type I patterns, analyzing the data from Experiment 2 (Type I) together with that from Experiment 6 (Types II and IV, same paradigm), the data from Experiment 3 (Type I) with that from Experiment 8 (Types II and IV, same paradigm), and the data from Experiment 5 (Type I) together with that from Experiment 9 (Types II and IV, same paradigm). In each case, a Firth-penalized logistic-regression model was fit with *Sought* as the dependent variable and with *Training Condition*, *Type*, and their interaction as predictors. Type I and Explicit-Promoting were the reference categories. In no case was there any significant effect of *Type* or interaction with it. The models were then refit, omitting *Training Condition*. The Seeking rate was significantly lower in the Type II condition of Experiment 9 than in the Type I baseline condition of Experiment 5 (by -0.66 logits, $p = 0.032$); otherwise, no significant or marginally-significant effects of Type were found. The predictions of the "sour grapes" alternative are thus not borne out.

14 General discussion

The results of the nine experiments are summarized above in Tables 3 and 4.

14.1 Algorithmic diversity in phonotactic learning

Participant behavior in phonotactic-learning experiments is more complex than has generally been assumed. Different participants approached the task in different ways, which we have identified with the implicit and explicit learning modes. Although the instructions and training procedure influenced how participants approached the learning problem, they did not determine it; Rule-Seekers and Non-Seekers were found in substantial numbers in every condition of every experiment. Naïve participants were able to consciously discover shared phonological properties and to invent mnemonic names for them to make them easier to think about.

In Experiments 6 and 7, differences in learning mode translated into significant differences in which pattern (biconditional Type II vs. family-resemblance Type IV) elicited better generalization performance (non-significant differences in the same direction were found in Experiments 8 and 9). These results suggest that there may be considerable uncontrolled between-participant variation in learning approach that can directly affect inductive biases. Distinct sub-populations may be using different learning algorithms whose effects dilute or even cancel each other.

Pattern type had little, if any, effect on report of rule-seeking (see discussion in Section 13). Methodologically, this is something of a relief, since it implies that within-experiment comparisons between pattern types are not likely to be contaminated by different choices of learning modes in the two conditions. Architecturally, it could mean that the two systems do not communicate with each other during problem-solving; i.e., that it is *not* the case that when the implicit system scents a pattern, it alerts the explicit system to search for a rule.

14.2 Phonotactic learning as concept learning

If laboratory phonotactic learning is a special case of concept learning, then previous research on concept learning provides theoretical and empirical reasons to expect Type II phonotactic patterns to be learned faster and better than Type IV, and for the Type II advantage to be magnified in explicit learners and reduced or reversed in implicit learners (Bradmetz and Mathy 2008; Feldman 2000, 2006; Kurtz et al. 2013; Lafond et al. 2007; Mathy and Bradmetz 2004; Nosofsky, Palmeri, and McKinley 1994; Shepard et al. 1961; Vigo 2009). Instead, Type IV patterns were learned faster and better than Type II (as in Moreton et al. 2017, Exp. 1), and explicit learning magnified the Type IV advantage in Experiments 6 and 7.

One response to these surprising results would be to conclude that laboratory phonotactic learning is *not* a special case of concept learning after all. Perhaps instead phonotactic learning in the lab is served by a different cognitive mechanism, with different inductive biases. However, to draw that conclusion from the

present results would be premature, since an alternative explanation is available.

The Type II advantage over Type IV in non-linguistic concept learning was found in low-dimensional spaces where every possible stimulus was presented in training. 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 others like voicing and vowel height that were randomized to make distinct stimuli (Section 3.1.1), as well as on visual features of the picture illustrating each word. They also varied in terms of ad-hoc phonological properties such as "ends with an [f]", which participants readily invented.

For an implicit parallel learner, irrelevant dimensions have little effect on difficulty, since the cue weights update in parallel regardless of how many there are. An explicit, serial learner in a phonotactic experiment, however, is faced with a difficult problem of identifying the pattern-relevant features amidst a background of distractor features, and if it does that by testing candidate features one at a time, the Type II pattern puts it at a serious disadvantage, as discussed in Section 13.

The results of Experiments 6– 9 could then be explained as follows: The implicit parallel system, being non-voluntary, is used by all participants, and learns Type IV better than Type II. (E.g., a generic gradient-ascent Maximum Entropy learner learns Type IV patterns faster than Type II, matching human performance on both phonological stimuli and visual analogues; Moreton et al. 2017.) The explicit serial system, impeded by distractor 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.[12] A prediction that follows is that reducing (increasing) the number of irrelevantly varied distractor features should reduce (increase) the Type IV advantage in both phonological and visual learning.

The present experiments thus provide no reason to think that phonotactic learning in the laboratory is anything but a special case of concept learning — a domain-general process (or processes) which only appears to be unique because the phonological stimulus space has properties rarely found elsewhere. It is a separate question as to whether the processes used in lab phonotactic learning are also involved in the natural acquisition of first- or second-language phonology (for a recent review, see Glewwe 2019, Section 1.1.2). The question has been approached from multiple directions, including neurophysiological similarities and differences (Domahs, Kehrein, Knaus, Wiese, & Schlesewsky, 2009; Hare, 2017; Moore-Cantwell et al., 2017; Wong, Ettlinger, & Zheng, 2013), and correlations between individual differences in lab learning and second-language learning (Ettlinger, Morgan-Short, Faretta-Stutenberg, & Wong, 2016). The most frequent

[12] Additional evidence for this account 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 11, 12, 13, and 14, 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 must have been either covertly using a more-complex rule than the one they stated, or they were using the stated one-feature approximation assisted by intuition.

approach, though, has been to compare inductive biases in the lab with asymmetries in natural-language typology, especially asymmetries in favor of phonetically-motivated patterns. The general picture that has emerged is that a pattern's difficulty in the lab is strongly influenced by its abstract structure, but only weakly, if at all, by its phonetic motivation (Glewwe 2019; Greenwood 2016; Moreton and Pater 2012a, 2012b; for opposing views on this contentious question, see, e.g., Finley 2017; Hayes and White 2013; Martin and Peperkamp 2020).

The strength of abstract structural biases and weakness of domain-specific ones strongly suggest that phonotactic learning, at least in the laboratory, is served by the same inductive processes that are used to learn analogous non-linguistic patterns. That at once raises the question of whether the numerous other biases seen in the acquisition and use of non-linguistic patterns (e.g., Kahneman 2011) also influence the acquisition and use of phonological ones. Do they show up in analogous phonotactic-learning experiments, affect natural first- or second-language acquisition, or leave their imprint on natural-language typology? These questions can only be answered by thoroughgoing comparative study of inductive biases in analogous problems across domains.

Acknowledgments

[Suppressed for anonymity.]

15 Appendix: Scoring guide for experiment questionnaires

Purpose: By following the instructions as literally as possible, any two scorers should (ideally) assign the same scores.

15.1 Did they mention Property *P*?

Purpose: This question is meant to find out whether they figured out each feature as a general property (not just as a list of sounds).

Do the following for each relevant property P:

1.1 Could anything ANYWHERE IN ANY OF THEIR ANSWERS plausibly be a lay-language expression of P?

(a) ANYWHERE IN ANY OF THEIR ANSWERS includes rules they said they tried and abandoned, statements that aren't rules, etc. — anything anywhere.

(b) Describing stress location in terms of schwa counts as describing stress, not as listing sounds.

If yes, score Mentioned P as TRUE and go to Item 2.1. **If no**:

1.2 Score Mentioned P as FALSE.

15.2 Did they state a rule at all?

Purpose: People sometimes check the boxes for "Tried to find a rule or pattern" and "Chose words that fit a rule or pattern", but then go on to state what actually a rule about *how they used their intuition* (e.g., "I went with what felt right", "I chose the one that was similar to what I had heard", "I chose the one that sounded more feminine", etc.). In this section, we verify whether people actually did report an explicit rule in terms of replicable properties.

2.1 Did ANY PART OF ANY ANSWER say or imply that their responses in A SET OF CASES were INFLUENCED by some PROPERTY P of the stimulus?

(a) ANY PART OF ANY ANSWER means we also count rules that they say they tried and later abandoned.

(b) SET OF CASES must be more than just mnemonics for specific words ("corn was something that made a popping sound when you pronounced it, like pop corn" does not count)

(c) A statement of *where* to look ("endings seemed to matter") that doesn't say what to look *for* does not count as a PROPERTY.

(d) Ignore hedging and epistemic or probabilistic qualifiers.

(e) Re INFLUENCED: It is not necessary to say explicitly which value of P went with which response. E.g., "I went by two versus three syllables" counts as TRUE.

If no, score Stated as FALSE and go to Item 3.1. **If yes**, continue to Item 2.2.

2.2 Was Property P scored as Mentioned P == TRUE in Part 1 1?

If yes, score Stated as TRUE, and go to Item 3.1. But **if no**, continue to Item 2.3

2.3 Was Property P specified so that the experimenter could replicate the participant's judgments as to which stimuli had or lacked it?

If no, score Stated as FALSE. But **if yes**, score Stated as TRUE.

15.3 How correct was the final rule?

Purpose: Here we score how much they figured out by the end of the experiment. We score each stated final rule as TRUE/FALSE for each of Correct, Approximate, and List, using the criteria below.

3.1 Was this participant scored TRUE for Stated in Section 2?

If no, score Correct and Approximate both as FALSE and go to Item 3.5. **Else:**

3.2 Make a "Scoreable Rule" by mentally editing their responses as follows:

(a) Ignore anything they said they tried but abandoned. If they didn't say they abandoned it, assume they didn't abandon it.

(b) Ignore statements that *only* give guidance for specific stimuli (e.g., "abupup was masculine"). (Do not ignore statements that use specific stimuli as examples to illustrate more general rules.)

(c) Ignore hedges and epistemic or probabilistic qualifiers.

(d) Combine all remaining rule-like statements into one big rule. E.g., if they say "F and V were feminine" and "B and P were feminine", score as if they had written "F, V, B, and P were feminine".

(e) Interpret statements of the form "stimuli with Property P are in Category A" as "*all* stimuli with Property P are in Category A".

E.g., interpret "disyllables were feminine" as "all disyllables are feminine".

(f) If they said their response was linked to a stimulus property, but didn't say how, interpret it in the way that makes it maximally correct. E.g., Interpret "I went by number of syllables" as "I judged all two-syllable words to be masculine and all three-syllable words to be feminine".

This procedure yields the "Scoreable Rule".

3.3 Is it clear that the "Scoreable Rule" gives a definite and correct answer on EVERY TRIAL?

If yes, score Correct == TRUE and Approximate == FALSE and go to Item 3.5. But **if no**, score Correct == FALSE and continue to Item 3.4.

3.4 Is it clear that the "Scoreable Rule" gives MORE THAN 50% correct answers on trials where it gives an answer at all? (It is assumed to give exactly 50% correct answers on trials where it gives no answer.)

If yes, score Approximate == TRUE. But **if no**, or if too complicated or vague to tell, score Approximate == FALSE.

3.5 Does ANY PART OF the Scoreable Rule list individual sounds, syllables, or letters?

If yes, score List == TRUE.

(a) ANY PART means List==TRUE even if the Scorable Rule *also* states a feature

(b) References to schwa, such as "starts with uh", counts as stating the feature "stress", rather than as listing sounds.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

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

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

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

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

Boersma, P., & Weenink, D. (2013). *PRAAT Version 5.3.14.* Software, www.praat.org.

Bower, G. H., & Trabasso, T. R. (1964). Concept identification. In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 32–74). Stanford, California: Stanford University Press.

Bradmetz, J., & Mathy, F. (2008). Response times seen as decompression times in boolean concept use. *Psychological Research*, *72*(2), 211–234.

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

Brown, J., Muir, A., Craig, K., & Anea, K. (2016). *A short grammar of Urama* (No. 32). Canberra, Australia: Australian National University. Retrieved from http://hdl.handle.net/1885/111328

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

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

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

Carpenter, A. C. (2010). A naturalness bias in learning stress. *Phonology*, *27*(3), 345–392.

Carpenter, A. 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*(2), 62–77.

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

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

Ciborowski, T., & Cole, M. (1973). A developmental and cross-cultural study of the influences of rule struc-

ture and problem composition on the learning of conceptual classifications. *Journal of Experimental Child Psychology*, *15*(2), 193–215.

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

Corbett, G. G. (1991). *Gender*. Cambridge, England: Cambridge University Press.

DeKeyser, R. (2003). The handbook of second language acquisition. In C. Doghty & M. Long (Eds.), (pp. 314–348). Oxford: Blackwell.

Do, Y., Zsiga, E., & Haverhill, J. (2016, January 7). *Naturalness and frequency in implicit phonological learning.* Slides from a talk at the 90th Annual Meeting of the Linguistic Society of America.

Domahs, U., Kehrein, W., Knaus, J., Wiese, R., & Schlesewsky, M. (2009, nov). Event-related potentials reflecting the processing of phonological constraint violations. *Language and Speech*, *52*(4), 415–435. Retrieved from http://dx.doi.org/10.1177/0023830909336581 doi: 10.1177/0023830909336581

Ọmọruyi, T. O. (1986). Adjectives and adjectivalization processes in Ẹdo. *Studies in African Linguistics*, *17*(3), 283–302.

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

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251.

Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M., & Wong, P. C. M. (2016). The relationship between artificial and second language learning. *Cognitive Science*, 822–847. doi: 10.1111/cogsc.12257

Evans, J. S. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *49*, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

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

Finley, S. (2017). Learning metathesis: evidence for syllable structure constraints. *Journal of Memory and Language*, *92*, 142–157. doi: 10.1016/j.jml.2016.06.005

Finley, S., & Badecker, W. (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* (pp. 706–711). Austin, Texas.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38.

Franco, K., Zenner, E., & Speelman, D. (2018). Let's agree to disagree: (variation in) the assignment of gender to nominal anglicisms in Dutch. *Journal of Germanic Linguistics*, *30*(1), 43–87.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=irr

(R package version 0.84.1)

Gerken, L., Quam, C., & Goffman, L. (2019). Adults fail to learn a type of linguistic pattern that is readily learned by infants. *Language Learning and Development*. doi: 10.1080/15475441.2019.1617149

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

Gordon, M. (2004). Syllable weight. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically-based phonology* (pp. 277–312). Cambridge, England: Cambridge University Press.

Green, P. S., & Hecht, K. (1992). Implicit and explicit grammar: An empirical study. *Applied Linguistics*, *13*(2), 168–184.

Greenwood, A. (2016). *An experimental investigation of phonetic naturalness* (Unpublished doctoral dissertation). University of California, Santa Cruz.

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

Hare, E. T. (2017). *Explicit learning of phonotactic patterns disrupts language learning* (Unpublished master's thesis). University of Massachusetts, Amherst.

Hayes, B. (1999). Phonetically driven phonology: the role of optimality in inductive grounding. In M. Darnell, E. Moravcsik, M. Noonan, F. Newmeyer, & K. Wheatly (Eds.), *Functionalism and formalism in linguistics* (Vol. 1: General Papers, pp. 243–285). Amsterdam: John Benjamins.

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic inquiry*, *44*(1), 45–75.

Hayes, B., Zuraw, K., Siptár, P., & Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language*, *85*(4), 822–863.

Haygood, R. C., & Bourne, L. E. (1965). Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review*, *72*(3), 175–195.

Heinze, G., & Ploner, M. (2018). logistf: Firth's bias-reduced logistic regression [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=logistf (R package version 1.23)

Høiland-Jørgensen, T., Ahlgren, B., Hurtig, P., & Brunstrom, A. (2016). Measuring latency variation in the Internet. In *Proceedings of the 12th international conference on emerging networking experiments and technologies* (pp. 473–480).

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

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Strauss, and Giraux.

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

Kelly, J. (1969). Vowel patterns in the Urhobo noun. *Journal of West African Languages*, *6*(1), 21–26.

Keren, G., & Schul, Y. (2009). Two is not always better than one: a critical evaluation of two-systems theories. *Perspectives on psychological science*, *4*(6), 533–550.

Kimper, W. (2016). Asymmetric generalisation of harmony triggers. In G. Hansson, A. Farris-Trimble, K. McMullin, & D. Pulleyblank (Eds.), *Proceedings of the 2015 Annual Meeting on Phonology.*

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

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

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

Kuo, L. (2009). The role of natural class features in the acquisition of phonotactic regularities. *Journal of psycholinguistic research*, *38*(2), 129–150.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (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*(2), 552–572.

Lafond, D., Lacouture, Y., & Mineau, G. (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, R. (2015). Learnable vs. unlearnable harmony patterns. *Linguistic Inquiry*, *46*(3), 425–451.

Lai, Y. R. (2012). *Domain specificity in learning phonology* (Unpublished doctoral dissertation). University of Delaware.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lee, Y. (1995). Effects of learning contexts on implicit and explicit learning. *Memory and Cognition*, *23*(6), 723–734.

Lewandowsky, S. (2011). Working memory capacity and categorization: individual differences and modelling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 720–738.

Lichtman, K. (2013). Developmental comparisons of implicit and explicit language learning. *Language Acquisition*, *20*(2), 93–108. doi: 10.1080/10489223.2013.766740

Lichtman, K. M. (2012). *Child-adult differences in implicit and explicit second language learning* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

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

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review*, *9*(4), 829–835.

Love, B. C., & Markman, A. B. (2003). The nonindependence of stimulus properties in human category learning. *Memory and Cognition*, *31*(5), 790–799.

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

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

Lumley, T., & Scott, A. (2017). Fitting regression models to survey data. *Statistical Science*, *32*(2), 265–278. doi: 10.1214/16-STS605

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

Martin, A., & Peperkamp, S. (2020). Phonetically natural rules benefit from a learning bias: a re-examination of vowel harmony and disharmony. *Phonology*, *37*(1), 65–90.

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: a synergistic effect. *Journal of Experimental Psychology*, *15*(6), 1083–1100.

Mathy, F., & Bradmetz, J. (2004). A theory of the graceful complexification of concepts and their learnability. *Current Psychology of Cognition/Cahiers de Psychologie Cognitive*, *22*(1), 41–82.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 355–368.

Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and non-rule-described categories: A comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1518–1533.

Moore-Cantwell, C., Pater, J., Staubs, R., Zobel, B., & Sanders, L. (2017). *Event-related potential evidence of abstract phonological learning in the laboratory.* MS, Department of Linguistics, University of Massachusetts, Amherst. ((Under review.))

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*(1), 83–127.

Moreton, E. (2012). Inter- and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, *67*(1), 165–183.

Moreton, E., & Pater, J. (2012a). Structure and substance in artificial-phonology learning: Part II, substance. *Language and Linguistics Compass*, *6*(11), 702–718.

Moreton, E., & Pater, J. (2012b). Structure and substance in artificial-phonology learning: Part I, structure.

Language and Linguistics Compass, *6*(11), 686–701.

Moreton, E., Pater, J., & Pertsova, K. (2015). Phonological concept learning. *Cognitive Science*. doi: 10.1111/cogs.12319

Moreton, E., Pater, J., & Pertsova, K. (2017). Phonological concept learning. *Cognitive Science*, *41*(1), 4–69. doi: 10.1111/cogs.12319

Moreton, E., & Pertsova, K. (2014). Pastry phonotactics: is phonological learning special? In H.-L. Huang, E. Poole, & A. Rysling (Eds.), *Proceedings of the 43rd Annual Meeting of the Northeast Linguistic Society, City University of New York.* Amherst, Massachusetts: Graduate Linguistics Students' Association.

Moreton, E., & Pertsova, K. (2016). Implicit and explicit processes in phonotactic learning. In TBA (Ed.), *Proceedings of the 40th Boston University Conference on Language Development* (p. TBA). Somerville, Mass.: Cascadilla.

Morris, P. E. (1981). The cognitive psychology of self-reports. In C. Antaki (Ed.), *The psychology of ordinary explanations of social behavior* (pp. 183–204). London: Academic Press.

Munoz, S. R., & Bangdiwala, S. I. (1997). Interpretation of kappa and B statistics measures of agreement. *Journal of Applied Statistics*, *24*(1), 105–112. doi: 10.1080/02664769723918

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

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

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

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Gauthier, P. (1994). Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, *22*(3), 352–369.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin and Review*, *3*(2), 222–226.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Onysko, A., Callies, M., & Ogiermann, E. (2013). Gender variation of anglicisms in German: the influence of cognitive factors and regional varieties. *Poznań Studies in Contemporary Linguistics*, *49*(1), 103–136. doi: 10.1515/psicl-2013-0004

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

11(6), 988–1010.

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

Pater, J., & Moreton, E. (2012). Structurally biased phonology: complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad*, *3*(2), 1–44.

Pater, J., & Staubs, R. (2013). *Feature economy and iterated grammar learning.* Slides from a presentation at the 21st Manchester Phonology Meeting.

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

Pycha, A., Nowak, P., Shin, E., & Shosted, R. (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)* (pp. 101–114).

Rabi, R., & Minda, J. P. (2016). Category learning in older adulthood: a study of the Shepard, Hovland, and Jenkins (1961) tasks. *Psychology and Aging*, *31*(2), 185–197. doi: 10.1037/pag0000071

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

Richtsmeier, P. T. (2011). Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology*, *2*(1), 157–183.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13, Whole No. 517).

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

Smith, J. D., Berg, M. E., Cook, R. G., Murphy, M. S., Crossley, M. J., Boomer, J., ... Grace, R. C. (2012). Implicit and explicit categorization: a tale of four species. *Neuroscience and Biobehavioral Reviews*, *36*(10), 2355–2369.

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in rhesus monkeys: a study of the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: General*, *133*(3), 398–404.

Smith, J. D., Zakrzewski, A., Herberger, E. R., Boomer, J., Roeder, J. L., Ashby, F. G., & Church, B. A. (2015). The time course of explicit and implicit categorization. *Attention, Perception, & Psychophysics*, *77*(7), 2476–2490. doi: 10.3758/s13414-015-0933-2

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

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgements

in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: a narrative review. *Applied Linguistics*, *36*(3), 326–344. doi: 10.1093/applin/amu076

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

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

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 908.

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

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

Wong, P. C. M., Ettlinger, M., & Zheng, J. (2013, may). Linguistic grammar learning and DRD2-TAQ-IA polymorphism. *PLoS ONE*, *8*(5), e64983. Retrieved from http://dx.doi.org/10.1371/journal.pone.0064983 doi: 10.1371/journal.pone.0064983

Zager, L. A., & Verghese, G. C. (2007). Caps and robbers: what can you expect? *The College Mathematics Journal*, *38*(3), 185–191.

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

Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*(1), 153–201.

Zubin, D. A., & Köpke, K. (1984). Sechs Prinzipien für die Genuszuweisung im Deutschen: ein Beitrag zur natürlichen Klassifikation. *Linguistische Berichte*, *93*, 26–50.